ABU DHABI – Domain Abuse Activity Reporting
Tuesday, October 31, 2017 – 15:15 to 16:45 GST
ICANN60 | Abu Dhabi, United Arab Emirates

DAVE PISCITELLO: …they derived that number is the number of customers time the number of mailboxes they pay for. That number is probably very low because you can also use the Spamhaus Domain Block List by doing a DNS query. So small businesses just have to do a DNS query each time they receive an e-mail to actually get the same information that others pay a fee for. There's a limit of about 300,000 queries a day on a small business, but that's certainly good enough for many companies that are small and medium businesses.

One of the other things that you can do and the article that we're publishing today actually shows is you can go to the DNS, you can look at Mail Exchange and Sender Policy Framework resource records. That would be the text messages of the major e-mail service providers and even some of the registries and registrars. You will see some of the RBLs that we list included in their SPF records.

Again, I think it would be very hard to go and say that people are not protected by RBLs, which is very important because that reinforces our assertion that we are showing you what the rest

of the world uses to filter and protect against abuse in the namespace.

This is a set of articles, a partial list of the citations that we use to help choose and prune down from the 86 possible lists of block lists that we knew to the ones that we actually use. There's one that I read recently that is really interesting and emphasizes how much of a threat spam is in a context we probably aren't very aware of but it's a very timely issue. The article is by [Vern Fuller] and it talks about how spam was used to influence political thinking and to provide fake political expression. This is something that certainly is a threat in my mind, certainly is affecting my country today.

These are the lists that we use. We only use domain block lists. We don't use IP block lists. We don't take an IP address and do a reverse lookup. We believe that's not sufficiently reliable. However, there are probably many domains that we don't catch as a result of this, so I would say that our numbers are possibly conservative.

Which leads me to answer the next question, do we capture all the abuse? We don't. Actually, none of the reputation block lists capture all the abuse. Part of the problem is that the way that most reputation block list operators capture abuse is by running what are called spam trap networks. They try to capture e-mail,

as an example. Then when they capture e-mail that should not have been delivered because there's no way in the world that anyone should actually know this e-mail address, hence the name "trap."

They all have different networks. They all cover different ISPs. They all have different parts of the world. No one could build a spam trap network that would cover all the providers. By using multiple providers, we hope that we have overlap, sure, but we hope also hope that we expand the set by union as opposed to intersection.

The most common comment that I receive about DAAR is, why did you choose spam? The GAC didn't actually say spam. Actually, the GAC did say spam. The GAC didn't say spam in their original Beijing Communique, but they did mention it in their Hyderabad correspondence.

Why did they do so? Partly is because they had conversations with people like me, and partly it was because they had conversations with others outside that community that explained that most spam are sent via illegal or duplicitous means. They've either used a stolen e-mail address, or they have built a botnet. The botnet is typically composed of computers that have been compromised. Breaking into someone's computer is a criminal offense in most countries. So these are

illegal. The machines that are sending spam almost always in some way represent an illegal act. If the spam itself is not an illegal act in some countries, it's certainly the case that breaking into the computers or stealing an identity is an act that covers many jurisdictions.

I mentioned before that spam is not just for e-mail anymore. It's actually a legitimate infrastructure. We see spam in link spam, spamdexing, tweet spam, messaging spam (SMS), and we're trying to gather information from all of those.

I mentioned that spam is a delivery vehicle. I'm going to publish an article later this week where I talk about spam. This is a separate article where I'm talking about spam as essentially being the same as a submarine force. It's a rather interesting military analog that I encourage you to read.

We mainly measure domain names that are found in the bodies of spam messages when we're talking about spam. The biggest thing that I want people to take away is that spam is important because this is what is in the lists that people use to block names.

Okay, now the exciting part for everyone. These are preliminary data. We feel very confident about them, but I wanted to characterize at a high level without naming names what we have seen so far in five months.

New TLDs, phishing is fairly low. It's lower than legacy TLDs over the five-month period. You see that there is some fluctuation. The same is true for malware domains in the new versus legacies. What we can basically characterize at a high level at this point is that phishing, malware, and botnet C&C activity is still largely conducted using names in the legacy TLD space. Where we see the biggest or shift in that orientation is in spam. Spam domains in the new TLDs are approximately the same rate as the legacy TLDs.

You do notice that spam domains took a sudden drop here. I'd like to think that was because everybody knows that we're building DAAR, but I'm not that egotistical. However, one of the things I will point out is that spam per month fluctuates.

For those of you who are familiar with a company called Cisco Systems, they have a webpage called the Talos E-mail and Web Traffic System. This is a product that used to be provided by a company called SenderBase. SenderBase was in the business of helping people who ran bulk commercial lists for profit to stay off block lists. They had a very aggressive manner in which they identified spam for their customers. So they could sell the service of, "We'll help you scrub your e-mail and make certain that it meets all the criteria so it's not blocked as spam."

I want to thank Alain Durand because he put in an enormous amount of time helping me do some visualizations. The circles represent top-level domains. As you can see, the X scale goes from 1,000 to 100 million. The center of the circle is the interesting part because that's what the score is. The size of the circle represents the number of registrations that resolve.

So you can see in phishing that there was one characteristic up here for two months. One TLD was a target of a concerted phishing effort. So the red ones represent a score that is at least ten times the average score. The score that I'm talking about is simply a percentage. It's the number of abuse domains over the number of registrations over a given day or in this case on the snapshot day. The last day of the month is the day that we chose for these reports.

Malware, again what we see is that there are only a handful of TLDs that actually seem to be targets for malware. Not surprisingly – [and it's surprisingly] – some of the legacy TLDs remain targets. There's quite a number of different theories for why legacy TLDs remain targets over new TLDs. One is that malware is often hosted on a compromised computer or a compromised server, and people who have compromised servers and those names have been around for a very long time represent a challenge in acquiring the right party and the right

instrument to go and suspend the domain or to remove content from that domain.

Botnets, again still we're only talking about six-tenths of a percent, so this is not a large number.

The interesting one is spam. If you see, there's a concentration of TLDs here that have remained persistently above the ten times average. Some of them were up in the 76-80%. Obviously, we're not naming names yet, but this is the characteristic.

Now this over here, just to make certain, we probably all know what this TLD is, but if you remember it's the center of the circle that's important and the center of that circle is below the average score. So there is a top-level domain out there with more than 100 million domains, and they do a pretty good job.

Showing the data in a little bit different fashion, I wanted first to talk about registration percentage. From May to September, what we observe is that the number of domains that were reported to DAAR from the gTLD zone data in the new TLD space represent between 11-12% of the total number of domains. That's 11-12% of approximately 190-195 million. Still the vast percentage of domains that are reported to DAAR lie in the legacy TLDs.

If you look at those that have at least one reported abuse domain – this is only those that have one reported abuse domain which typically is between 350-380 TLDs each month – approximately 44-49% of the abuse domains are reported in the new TLD space. So that means that 11-12% of the resolving names in our data have approximately 44-49% of the malicious registrations.

That number, everybody goes running to CircleID or to Domain Incite, and they're going to go and yell, "Oh, my God! Oh, my God! The new TLD space is just horrible," right? You can't just look at data 100,000 feet and come up with that kind of answer. The value of DAAR is that what we can do is go and carefully look and sort and discover that in May the five most exploited new TLDs represented 22% of the resolving names but 56% of the abuse domains. If you take out the top 25 most exploited TLDs, they represent 70% of the resolving names but 97% of the overall abuse registrations that are in the new TLD space for each of those months. I just show you May and September because, size.

If you notice, there's a little bit of change but not a great deal. So we have a persistent set of outliers that statistically are interesting to us and the question is, what do we do with those? How do we use this data?

The next build of DAAR will hopefully have reliable registrar data. So the next effort that we want to make is, which registrars are actually sponsoring the domain names in these 5, 10, and 25?

One of the challenges we have had so far with constructing registrar data with accuracy is collecting WHOIS. How many people collect WHOIS in volume? What's the biggest problem that you have? The biggest problem is that registries rate limit on an IP basis. For us, this is a case of trying to build the system the way that others have to build systems so that we can claim that you can reproduce our methodology. So we do the same WHOIS querying through the public WHOIS system that everyone else does. We have finally managed to grow our infrastructure to do this so that we can keep up with new registrations.

One of the things that we are struggling with is that in order to be able to make certain that we attribute a domain name accurately and reflect things like inter-registrar transfers, we need to keep some historical WHOIS for each of these names.

So where are we? Why is this the limit of the data that we are sharing right now? I'm actually channeling our CEO Göran Marby and my CTO David Conrad who essentially say doing it right is more important than doing it fast. I could give you data and if someone in the community was really unhappy with that data, we would look like we're not good scientists, we're not good

engineers. There also may be questions of whether or not we've actually used the data that we are licensed to use in the manner that was intended.

So we have two efforts going on. One is to review the data feeds and licensing with our developer who is reviewing it with the people for whom we have contracted and subscribed data.

The second is that we have contacted and have a commitment by two people, one in academia who is very well known and one a very respected member of the private security and operations community, to review our methodology. They're going to take our documentation. They're going to look at it and they're going to say, we hope, "Yes, we think that what you do is correct. These are the things that we like. These are the things that we don't like." Anything they say that they don't like, we are going to fix. Our goal is that independent review will validate what we do as meeting industry practices.

I mentioned earlier that we're trying to tune our collection systems to ensure resilient operations. I also mentioned earlier at the ccNSO members meeting that we have contracted with our developer for a new set of features for Version 2. Among those features is some additional automation for reporting. Now that we have an idea of the kinds of reports that we like to have prepared for us out of the database, they are building in a

mechanism to be able to actually pull that down on a daily basis.

We also are looking into granular attribution. One of the things that we do right now is we have a composite identification of the attribution, which block list gave us this domain. Alain actually made this recommendation that it would be helpful for us to be able to say, "Okay, Spamhaus, SURBL, and SpamCop all identified this domain as abusive. In this second domain, only Spamhaus did. This third domain, only SURBL and SpamCop did." And have that help us understand how much trueness in addition to precision we have.

We're also experimenting with some additional measurements. Because it's experimental, I'm going to just not mention them.

I'm going to finish with Dave's Wishlist. I think I spend way too much time looking at this right now. I should probably take some vacation. The industry at large is not really good at discriminating between legitimate and compromised domains. That's a second-order review.

The fact is that whether it's legitimate or compromised, it doesn't change the matter of whether the domain or URLs at the domain represent a security threat because if I registered a domain ebayloginfreeprize.com and that site hosted an eBay impersonation login page, it's a security threat. If I registered a

domain legitimately for marysembroidery.com and that page had a URL that said marysembroidery.com/ebay/login.php, it's a security threat.

So we look at legitimate versus compromised at this time as a composite. It doesn't matter; they're both threats. We're going to try to encourage the community to do better. John and I and Carlos and others in our company are members in good standing in the Anti-Phishing Working Group, MAAWG, other places where people who generate block lists participate. We're pushing, and they're actually really excited that we're doing this and are listening.

Also, I'd like to do tracking of recidivism. There are some registrars who will take down a domain, but once they've taken down a domain they put it back into the available pool. And either the same criminal or someone else will register that domain almost immediately after it's dropped and it becomes live again. This creates a lot of challenges for the RBL providers because they have to be very careful. Is my strategy to eliminate the domain from my list when it doesn't appear in the zone, or do I pay attention to the fact that there's recidivism and keep that domain in my RBL for a longer period of time? So this is another tack for our analysis.

I'd like to also start to find other classifications of spam domains. We already know that we're looking at things like botnets, malware, C2. What about counterfeit goods? What about 419 advance fee fraud? Other kinds of manipulations and social engineering, can we characterize those? I've talked to a couple of universities and they have some ideas. They also have what I don't. They have lots of really smart people who are dying to write a Ph.D. on something that would be useful. So I'm really keen to try to get some of those people onboard with doing some analyses. Some of them already have labs where they collect phishing data. They have their resources. They have the e-mails. They have the malware hashes. And so I'd like to do that.

Speaking of malware hashes, I'd actually like to start to think about what domains are hosting what kinds of malware. So that's another place to go. And I won't stop.

What about URL amplification? Some spammers will register one domain and they'll create many, many URLs for different attacks: marysembroidery.com/ebay/login.php, marysembroidery.com/hsbc/phishinglogin.html. You get the idea. Sometimes there are amplifications in the hundreds or thousands on a given URL.

Other spammers will use subdomains because often we users will stop reading a URL as soon as they come across the name that they're looking for. If I have a domain that I want to look like apple.com, I could use appleguideance.com, which is not registered by Apple but because the first thing the e-mail recipient sees is "apple," especially those who are not familiar with what a URL is at all, they're going to click on that.

So using subdomains or URLs has a very interesting amplification for spammers and phishers. I'd love to be able to start using our data, and we have all those URLs, to try to see if that's interesting.

Another thing that I want to pivot on is IP and autonomous system number reputation data. I see John raising his eyes going, "Dave, get a life." But these are things that others are doing. There's a really good product out there, a commercial offering, called Seclytics. I know the guy that built it, and I want to go talk to him and say, "We'd like to figure out how we can incorporate your data which identifies IP reputation and ASN reputation with our data and what do we come up with." That's just more interesting stuff.

Registration fees, everybody in this community makes accusations or allegations that inexpensive registrations attract

spammers. Well, who has the data? Nobody has the data. We should have the data.

And then as I said before, we're literally turning away offers for reputation data. Part of it is to try to have the same degree of confidence and universal acceptance, so to speak, of the reputation data that we are going to receive.

So that's it. I'm out of breath. If you want to talk to me, if you want to scream at me, certainly you can try any of these. Or you can ask some questions now. Thank you very much for coming, and I hope that this project is as interesting to you as it is to me.

UNIDENTIFIED MALE: Okay, let's take some questions while we still have time. I know we had a question over here. We have the gentleman down here.

DIETMAR LENDEN: My name is Dietmar Lenden, Valideus. I've just got a question about Slide 18 where you make a statement at the top of the slide, "All gTLDs having at least 1 reported abuse domain." I'm curious about that, and I'm assuming that "gTLDs" is also referring to new gTLDs, so legacy and new gTLDs.

DAVE PISCITELLO: Yes.

DIETMAR LENDEN: So we've got brand TLDs that have a single name.

DAVE PISCITELLO: No, no. For example, let's say we're looking at May. In May, I believe we had 365 TLDs new and legacy with total of one or more abuse domains. So there were 900+ domains that had none. They're not depicted here. They're actually calculated in the abuse score because they can get one at any given time. Well, maybe they can't because some of them aren't going to resolve any names. But this only represents the subset of new and legacy TLDs that have at least one abuse score. The same is true for IDN gTLDs.

Other question? Name?

SIMON JOHNSON: Simon Johnson. I'm the Board Director of auDA. The questions that I've got are probably more of an observations as well. Probably in the interest of disclosure, I actually created a company that did exactly this back in 2013. So I certainly share and understand your passion for this sort of stuff, and it's really exciting to see.

Two questions. Where does ICANN draw the line in developing a product, per se, as opposed to leaving this to the private sector to deal with?

DAVE PISCITELLO: That's above my paygrade. Actually, I don't consider this a product. This is a project.

SIMON JOHNSON: Project, not product.

DAVE PISCITELLO: OCTO was given a directive by the CEO to go out and do research that would help us provide information or data to support policy development or to support a consideration of how well currently policy is performing. So that's really what we did. We're not going to sell this. I don't believe we're ever going to sell it. We're probably going to ask for money to continue to support it, but my understanding is that this is going to remain a system that we will use similar to the other reporting systems that ICANN has for registrations and the like.

UNIDENTIFIED MALE: So in terms of what success looks like, it's decision making. When you work for a nonprofit organization like ICANN, and this

is not the first one that I've worked [from], you always have to consider the effect you're having on the industry out there. But you should also not let that freeze you from doing anything.

So one of the purposes of this system that is now called DAAR – it has had other names – was actually to drive the policy discussions. We've had a lot of discussion inside ICANN about what is abuse and what are we doing about abuse. What we were really trying to do, apart from the fact that we're geeks and we love the interesting data, was actually to have this – this kind of discussion. So it really is a tool to help us shed light on what's happening out there so that we can have these kinds of discussions about what we should and shouldn't be doing.

Dave's right. I don't think this would ever in any way be able to become a commercial product, and we hope that we're not competing with other commercial products. We're trying to drive a policy discussion, not necessarily provide a tool that tackles the exact abuse issues.

DAVE PISCITELLO: Let me just add one thing to that. One of the reasons why we are so keen to make certain that our methodology is transparent and reproduceable is because if someone else wants to go build this, we're delighted. Quite honestly, it's a 24-hour job on top of

my other 24-hour jobs, and my wife is threatening to leave me if I don't stop talking about this at 4:00 in the morning.

But yesterday during the Cross-Constituency Working Group, somebody asked, "Why should ICANN do this?" From I believe one of the Government Advisory Committee there was an expression that ICANN ought to be doing this as part of its public interest responsibility, that having third-parties do it doesn't have the same gravitas in some respects as having us report on behalf of the community.

Go ahead, Paul.

UNIDENTIFIED MALE: It was mainly an observation. I [notice] just from my experience in discovering this system, one thing you said that interests me was that you mainly track domains and not IP addresses. I found that to be really interesting because spammers don't necessarily register a domain name. They use a botnet. They will go and sign up for a server, a VPS somewhere. They'll send a lot of spam. I think that's reflected in your graphs as to why the botnets are so low. Is that an area that you're going to focus on? Because I see it as a pretty big – I won't say flaw – but it's a [hole there].

DAVE PISCITELLO: It's a great question, and I'll give you a couple of answers. One of the things that, obviously, we were interested in is the domain name space. When we get an IP block list and we experimented with this, doing reverse lookup is unreliable because the spammers don't actually put anything in ARPA. So that was one consideration. We could do more to get that, we know that. But in order to get the project at level one, crawl, we did not include that.

The second is that we actually were concerned that there would be pushback from RIRs or concern from RIRs that we were doing something that they should be doing. That didn't carry a lot of weight because I don't really care. In this point, I'm a scientist, not a politician. So I said, "No, it's important."

What we do try to recognize is that a lot of botnets do algorithmically generate names. They're reflected here. We have Bam and xBot and BotList and things like that, that feed into what we do. So we're trying to get as much of the domain-related reputation.

As I mentioned earlier, I think these are smaller than the actual numbers of origins of spam. The other thing that I mentioned is that by and large from e-mail measurements, we're really looking mostly at URLs that appear in messages or in attachments, not senders. You could do a different study that

focused on senders, and you're absolutely right. IP senders would be very significant there.

It's a great place to go. What we're going to have to do is figure out how we can do that well. I want to be able to do this really well and then say, "Okay, now I have a baseline for how well it has to be for us to get in IP addressing." Does that answer your question? Okay, great.

UNIDENTIFIED FEMALE: Dave, I have a couple of questions from the chatroom.

DAVE PISCITELLO: Okay.

UNIDENTIFIED FEMALE: From Matthias Pfeifer from .berlin, "Will the collected raw data the DAAR is using for the statistics be available to ROs so they can use the data for their own anti-abuse solutions or actions?"

UNIDENTIFIED MALE: One of the reasons we published the list of who we use is they could go and do that. Our licensing agreements don't allow us to pass through the raw data. The reason we are being so transparent and we try to use publicly or commercially available

data is so that others can go and do this if they want. But our licensing agreement does not allow us to pass through raw data.

UNIDENTIFIED MALE: There's another possibility. That possibility would be that the community would say, "This is what we want ICANN to instrument." And then we would go off and we would negotiate with the reputation block lists for a fee that would allow us to pass through that information. The current license we have says that we can share derivative data – aggregated numbers and things like that – but we can't take a sub list that identifies the abuse in .berlin and give it to .berlin under the current licensing.

Anything is possible. As everybody in this community knows, it's a matter of money. If that's what the community wants, we will go and we'll negotiate. Building the API for that is not going to be rocket science. Figuring out the subscription model, figuring out how we would pay for it is probably the 90% effort.

UNIDENTIFIED FEMALE: Another question from Matthias, "All TLDs have at least one reported abuse domain. Does that include all the brand TLDs?"

DAVE PISCITELLO: No. I think I just answered that. What this says is "all TLDs that have" at least one reported abuse, not "all TLDs have" at least one abuse domain. Maybe the language is not clear. For example, in this slide as I mentioned earlier, in May we had 356 TLDs in our portfolio out of the 1240 that had one or more and that's what's represented here.

REG LEVY: Reg Levy from Tucows. Looking at this slide, is this the manner in which you plan to report your findings in the future? You said we're not naming names yet, so that implies that you're planning on naming names. This is an aggregate of what it looks like in gTLDs, but are you planning on calling out gTLDs and registrars in the future?

DAVE PISCITELLO: You can take that.

UNIDENTIFIED MALE: So the "yet" is we don't know. We have some people in the community who really want us to publish with names and we have other people who say that it shouldn't be published with domains. I'm not a lawyer in this life or any other, so we're having those discussions. A lot of this is and should be driven by the community. These are the kinds of discussions we're

actually happy to have and one of the reasons we did this is, "Well, let's have that discussion." What level of transparency should be on this? What level of, I guess, risk is ICANN as an organization going to take on this? And that's partly or I would say for the majority a discussion to have with the community.

DAVE PISCITELLO: One of the things I want to point out since I know this data probably better anyone but Greg Aaron is that there's a very small number of outliers. An overwhelming number of contracted parties who come to ICANN meetings are below that norm. I will talk to you personally about how great you guys are.

REG LEVY: I'd like to ask a follow up with regard to how the data is presented. Are you thinking that this is going to be shared internally at ICANN with names named? Would this be the basis for Compliance to raise any issues with either registries or registrars?

UNIDENTIFIED MALE: That's another interesting discussion that we're having with the community. We've looked at some of the underlying data to see if there are ways to use this and how we would use this. Jamie

Hedlund leads that side of the organization, and I talk to him on an almost daily basis. The answer is we don't really know yet.

Obviously, when we see outliers – and when I say outliers, I mean outliers – data often doesn't give you answers. It gives you questions. So when we see those, it's going to give us base to go and ask questions of ourselves and ask, "What's causing this?" It's not yet clear that the underlying data, the raw data that we have, is something that we can use in those realms as a compliance tool. That's a discussion that we're having internally and, I think, a discussion that we're having with the community.

I wish I could give you a straight answer, but I don't think there actually is one at this moment. I would love a scenario where if there was an outlier in the industry that was causing issues not for ICANN but for the whole industry that we could say, "Well, you're in the system and therefore you're bad." But the world doesn't work that way and you have to be very careful when you're looking at data to actually understand what the underlying causes are. So, yes, when we see outliers, we sit there and we go, "Well, data gives you these questions," and we're trying to answer them. But DAAR system is just one part of that.

There is so much behind this data that we are constantly looking at and trying to figure out what this means. And we do talk to Jamie and the Compliance people and say, "Well, do we know

why? Is there something we can do? [Are there are other ways] that from a Compliance perspective, which is not our area of expertise, you would like to look at data [or are there] better tools that you need?" Because maybe DAAR is the wrong tool for Compliance. These are all good conversations to have both internally and more importantly with the community.

As far as the reports go, we would like to be as transparent as possible and publish as much as we can to the community. So if there's any report [that's inside], we would love to be able to make it outside within the limits of our licensing. So transparency is the goal always.

DAVE PISCITELLO: I hate to answer your question with a question but, as an operator, if we saw something anomalous, would you want us to talk to you about it?

REG LEVY: I'd want you to come to me directly.

DAVE PISCITELLO: That's fine. I'm looking for those kinds of answers. That's fine.

REG LEVY: And I'd want you to do it and not have it as a Compliance [complaint].

DAVE PISCITELLO: Okay, we can notice that. What if it's not you and it's somebody who is an outlier and is completely unresponsive? Our exchange is what we're struggling with.

REG LEVY: Then you start with the "me" there, and then it [inaudible] up to it.

DAVE PISCITELLO: Yeah, and I think one of the things that's real fascinating is that I know what's behind the curtain. I've also had plenty of conversations with people in the registry and registrar world. I've actually had them come to me and go, "This player over here, I really hate them." There's an opportunity for the good practicing members of both stakeholder groups to sit and be thoughtful about how we can make certain that everybody is self-regulating. Because it's the people who are in the self-regulating world, who have chosen not to, who are really contributing 95-97% of the problem. That's like identifying a part of your population that really shouldn't be getting on

subways with the rest of the population. How do you fix that? I have ideas, but….

SIMON JOHNSON: I just wanted to follow up a statement you that you made before. Just for the record, I'm a director of auDA and a board member and chair of the Security and Risk committee. I wouldn't mind you guys just factoring into your thinking entities like [us] who are a little bit different from other CCs in that we're a regulator. And as a director, I'm personally liable for .au, and the government discharged its powers to us under a telecommunications act. So when it comes to this sort of stuff, security, I would want you guys to come to me and say, "Hey, we have a problem," and give us the opportunity to deal with it rather than push it out there and blow it up. I just wanted to put that out there [as] a discussion you guys want to have.

UNIDENTIFIED MALE: Were you just offering us your data?

SIMON JOHNSON: No.

UNIDENTIFIED MALE: Okay. Our goal is to solve the problems. As operators, this is a scientific, experimental thing. So even before we had this, whenever we've seen issues, always the best approach has been to go and talk to the people with the problem privately and say, "Hey, are you even aware that you have this issue going on, and can we as individuals help?" So the idea is not to go and name-and-shame people or to cause people problems. The idea here is to shed light on the data and then to work with the community to see where the solution space is. Some of it is probably in the policy realm. It may be in the compliance realm. I don't know. But at least we're having the conversation.

SIMON JOHNSON: Yeah, and anything we can do to make this a safer zone we'll absolutely do it, and we're a pretty safe restricted zone compared to the rest of the world.

DAVE PISCITELLO: I think the conversation can begin both ways. If you want to get on the phone with [Graeme] and we could talk about what we see, happy to do that. I don't think Göran or David is going to object to that. Same with you. Because I think the more that you guys see our data and get a sense of what we're doing, the more comfortable you are. And the ones who become comfortable,

you word of mouth to the rest of the community and you assuage fears. So I think that's really a valuable opportunity.

But the other thing I want to point out is that this kind of reporting is what you see from other people than ICANN. That granularity paints a very negative picture of the new TLD space. This does not. This helps everyone – Government Advisory Committee, every one of the other constituencies – to see that problems that you guys already know. You're in the business. You know who these people are. I probably don't have to tell you. And you also know the registrars.

So you tell me. What's the best way to make certain that people are looking at this and not at that? Especially the registrars. I'm sure that you register new TLDs. This is not what you want people to see. The only reason I put this up is the start reminder that this is what's going to play on the sites, not the details.

[SIMON JOHNSON]: Could I give you an answer to that just from our perspective? And I'll say from the outset it's very different from how the U.S. and others operate. If we found out that there was a phishing/malware/bot server with an .au domain – I don't recall it ever happening, but if we did – certainly we could lock it, we could drop it into pending delete. We don't need a court order or anything else to do that.

U.S. is a very different place because I know if something was to happen, it's me on the line. So they're the sorts of consequences that come from this data, so absolutely we would want to know and we would do something about it straightaway.

UNIDENTIFIED MALE: I think you did. I think I just found one of your guys, and they did something.

DAVE PISCITELLO: Okay, I'll give you all two minutes of your lives back. Oh, I'm sorry. One more. I apologize.

KAL FEHER: Kal Feher, Neustar. I want to understand how you account for false positives.

DAVE PISCITELLO: This is a longer conversation than two minutes. I just had a Skype exchange. There are actually, I think, at least three perceptions of false positives. One is the extent to which one of the reputation block list providers that we use exhibits a false positive rate. Most of the providers that we choose have false positive rates of less than 4%. Typically 2%. In fact, I believe all of them do.

The second is distinguishing a complaint that a registrar or a registry receives that says, “This is a false positive.” My experience is that often the people who are making complaints have actually violated the policy and sometimes unwittingly.

Because you go anywhere on the Internet and you search for “why am I block listed?” you go and you find 1,000 articles that say you’re block listed because you bought a mail list and that mail list had a domain that has been added to a block list on it and you sent mail to a spam trap and that’s why you’re block listed. People who don’t quite understand the whole mechanism come running to Tucows or running to any of the registries and registrars going, “I shouldn’t be on this block list. You shouldn’t have taken down my domain.” Then if you go and you look at what they’ve done and you dig down, you’re on a block list.

There’s a little-known fact: ICANN was on a block list last year during one of the ICANN meetings because somebody went to one of our wikis, signed up as a an ICANN community member using a mail address that was on a block list. We didn’t filter properly; we got our hand slapped. We went to SpamCop. We talked to them. They said, “Gee, we’re really sorry, but this is the process.” So we ate the dog food. We went and we completed the process, and they took us off the block list.

There are so many ways that there is misinterpretation of how this works. It's just really frustrating.

KAL FEHER: If I could just follow up on that, it really doesn't matter for an organization if you have a false positive for one of these lists because they're still very effective with the odd false positive. You reverse that though, you use this data for the product you were talking about earlier, the Internet police thing, a false positive might take someone's domain name off them. So even if it's a low percentage….

UNIDENTIFIED MALE: Let me talk to that. As you know, I work on a lot of these cases with law enforcement and stuff and I'm very much – in fact, they hate me for it a lot of the time – I'm always about the unforeseen consequences of thinking action and [hurting people]. Every one of these names that we're saying is on these lists, is on these lists.

Now, this is not data that we're saying people need to take this data and just go take action. Which is part of the discussion about compliance: can we use this for compliance? That's a really interesting discussion because we know there are some

false positives, so how do you weed them out? Those are discussions we have to go have.

But don't get the perception that we're saying people should just take this and that we're giving them a [real feed] and they should then just go and act on that blindly.

KAL FEHER: That's not what you're saying, but I think that's the result that's going to happen.

UNIDENTIFIED MALE: Well, let's work as a community to make sure that's not what happens.

DAVE PISCITELLO: Yesterday I had this exact same conversation with [Alan Wood] and he actually agreed with me because I said that this list is not something that I think anyone should hand to a registry or registrar and say, "Take all these domains and suspend them." If you take a name off this list and you don't go and put the effort in to find the e-mail, visit the website, and do all the normal investigation that you probably already do, then you're only doing half the job. We're not doing the work for you. We're

helping you by gathering the lists themselves, if we ever are allowed to share them.

So at this level, I think our conversation isn't particularly important. If we go to the next level and there is some way for us to pass along data, we've talked about, "Well, can we pass along other data?" Can we go out, for example, and on a phishing domain curl the page, zip the page, provide that with you? Maybe we can do that. If the community is willing to spend money for us to do that, then we could do that because it's like another terabyte of storage. How bad is it? It's a little bit more CPU.

Look at me: "Just give me some more money. We can do that for you."

UNIDENTIFIED MALE: I'm going to put you in a room with the lawyers.

DAVE PISCITELLO: Yeah, I know. I know. Just don't put me in a room with the lawyers too long or I'm gone.

But thank you. We should talk about it more as we go further down because I agree with you. I'm not interested in making anybody look bad. I'm interested in eventually having nothing to

talk about. That they found somewhere else to go and do their malicious stuff and it's not my DNS, quite honestly. It's my DNS, by the way.

UNIDENTIFIED MALE: Could I just ask one other question? Just following on from that, something I did in another life was an API, so I can see the next evolution of that combined with the gentleman's question at the back is someone is going to come and say to you, a registrar or somebody, "We want an API. We want to start querying." I think that's potentially where the issue would come from.

The second part of that is really data currency. I know from doing lots of WHOIS queries and IP reputation queries and things, it's this domain was affected at a point in time. And I think when you're talking to people about it, it's always important to say, "Hey, this was current as of blah, but it may not be now." So if you're relying on this data to go and scrub lists with, that's when you run into problems and the appearance of false positives might come up.

UNIDENTIFIED MALE: Yeah, absolutely. The questions of APIs, I don't see that happening any time in the near future or pretty much in any future because then we come back to the original question that

you asked which is about, is this ICANN's role? And there is an industry out there that does this kind of thing. I suspect there would be a lot of discussion about having a nonprofit do this stuff. But that's for the lawyers.

And, yes, anybody who takes these kind of feeds and doesn't do secondary analysis on the data and look out for the unintended consequences, thinking down some poor grandma's website that she actually relies on for her knitting industry – I'm only half joking there. The most important name in the world is always yours. When you take this kind of data whether you're looking at DAAR or any of these other feeds, I'm expecting if you have a group inside your organization doing this operational security work, that they do their due diligence.

That's not what DAAR is. DAAR is giving a window onto the data that's out there and that people are using. It's not, in my opinion, fit for the purpose on its own or even if somebody was to build this on their own, there's a lot more that goes into deciding whether or not you need to take action on a name. This is not actionable data as it stands.

DAVE PISCITELLO: One of the other things I'd ask people to be thoughtful about is there is a fair bit of acrimony between reputation block list providers and hosting companies – DNS hosters, registrars,

registries. One of the things I would really like is for data like this to create an opportunity for those of us who work with them to create a positive environment for conversation. Create an environment where the RBL people, for example, start to appreciate let's provide a little bit more data directly to, in addition to.

I find, honestly, that lots of the reputation block lists don't have a lot of historical data. Part of it is that some of them don't have the financial backing to actually have petabytes of data and store it with a fair amount of permanence. Others are restricted by the kind of data that they have to a limited amount of time that they actually can hold it because they are holding some information like e-mail messages that have personal data.

So I think if we all can have a conversation of, "Well, what would you need? What do I have? What should I keep? Can I keep it? Can I share it?" especially in the GDPR context, maybe that's the next step in trying to assimilate the industries instead of being combative, cooperative. Then there's a benefit for all the good players on both sides.

UNIDENTIFIED MALE: If I could just quickly respond, I think the commercial realities mean that if you release this data for free, people will use it and they're not guaranteed to add additional processing. You might

like them to do it and we might encourage them to do it, but the reality is that if you can get away with using it as is, you probably will.

And these feeds can be very effective for someone with a reasonably high false positive rate because the normal target market a false positive doesn't bring down their systems. They might miss the odd important e-mail, but there's no strong commercial demand going the other way for these feeds to improve their false positive rates.

So we're looking at feeds here that there's a level of false positive that's just fine for their consumer base and they'll never need to get better than that because their consumer base will never be harmed by having it at that rate.

DAVE PISCITELLO: I think I'll push back just a little bit because the people that we use and the people that we talk to on a regular basis do believe that if their false positive rate became suspect, that they would lose customers. So they have the same worry about their reputation as anyone who has a domain name.

As an example, I've known the Spamhaus people for well over a decade and I've worked with them for a very long time. If I go

and talk to them about somebody who comes to complain with me, I get a very clear answer about what they did.

I'll give you a case in point. China Guangdong network was taken down. The entire network, the entire IP space, over a million IP addresses could not send mail using Port 25, and everybody went berserk. How could they take down a whole IP space of a million mailers? The reason why was because they have 1,600 unanswered reports. They just refused to engage with Spamhaus and so Spamhaus said, "Okay, if the carrot didn't work, we're going to use the stick." Four hours later, all 1,600 complaints were corrected. So was that a bad thing or was that a good thing?

UNIDENTIFIED MALE: That's a bad thing, bearing in mind that it's not a harm in the context of what you're providing. But if you use that in the reverse, you're taking domain names off other people simply by association. Then you're thinking legitimate domain names that may not have been participating in spam at all.

DAVE PISCITELLO: I think we can have another conversation about this and I can show you some IPASN data. I had this conversation with Paul. DigitalOcean is full of IP addresses that are block listed. Should

we actually turn off access to DigitalOcean and not route to them because they have this? That's pretty dramatic. But what else do you do if they just don't keep up and they just aren't responsive? This is the meta discussion of what else do you do when somebody is in a self-regulating industry and refuses to abide by the rules.

UNIDENTIFIED MALE: Well, I think you've touched on the problem here, that it's commercially challenging. It's a lot of effort.

DAVE PISCITELLO: Yep.

UNIDENTIFIED MALE: So the simple alternative is to block the space, but the more challenging alternative is to do it almost incident by incident, which is why I fear that if you release this for free, people won't go for the challenging alternative. They'll go for the simple one, which means we'll have a lot of domains that will go dark for no good reason.

UNIDENTIFIED MALE: I think that's very good input, and I don't think there is any intention, certainly not at this time, to give this data out for free.

We're very aware of these issues. I think if that discussion does come up in the community, because that's where it will come from, I think it would be very good for you to get up to the microphone and say, "Beware what you do" because it's a real risk.

I think we're going to close up the microphone here, but I think these kinds of conversations are exactly what we need to have. So I'd like to thank everybody for coming here and raising questions. I'd like to continue having this discussion as we develop the system further. As we get to the stage where we're thinking about what we're going to report, I'd like to have discussion the community about: what are the things we should be worrying about? Are we doing the right thing? Is there something else?

When we have these kinds of discussions, I'd really like that discussion not to be between me or Dave and the community. I'd like the community to be having that discussion. We're very biased in the fact that we've been developing this system. I'd like the community to have that discussion and say, "What are the goods and the negatives and the dangers, and where should ICANN be taking this?" rather than having staff drive it. Obviously, we'll end up doing all the work that you make us do, but that's life. That's what we're paid for. But I want the

community to keep having this discussion and to give us feedback.

If there are other things that you think that we should be measuring, let us know. If there are things that you think we should be very wary of, be vocal. Let us know so we can take that in. And believe me, we have lots of lawyers and folks like that who give us very good advice about exactly stuff like what was raised about the risks.

So I think with that, we're pretty much at time. Thank you, Dave. Thank you, everyone.

DAVE PISCITELLO: Thank you all so much for coming. I'm really excited that you're excited.

[END OF TRANSCRIPTION]