COPENHAGEN – ICANN GDD: Statistical Analysis of DNS Abuse in gTLDs Study Results Preview
Tuesday, March 14, 2017 – 11:00 to 12:15 CET
ICANN58 | Copenhagen, Denmark

BRIAN AITCHISON: Okay. Thanks very much, everyone. Welcome to the Statistical Analysis of DNS Abuse in gLTDs Study Preview. My name is Brian Aitchison. To my right sits Drew Bagley from the Competition, Consumer Trust, and Consumer Choice Review Team. Our researchers from SIDN and Delft University of Technology include Maciej Korczynski from TU Delft and Maarten Wullink from SIDN Labs.

We're going to walk you through how far we've gotten on this study so far, and we're going to talk about its background and methodology and take your Q&A on that.

Next slide, please – there you go. There you have our agenda. Next slide, please. I'll go through the study background. I think there's – ah, right.

So it has quite a lineage. Back in 2009, prior to the expansion of the DNS through the New gTLD Program, there was a question posed to the cybersecurity community that is encapsulated in this memo that you see hyperlinked at the top: "Mitigating Malicious Conduct: New gTLD Program Explanatory Memorandum." Essentially, ICANN asked the security

community, "What can we do to mitigate malicious conduct issues should the DNS be expanded through a new gTLD program?"

The questions that the community came up with you can see displayed in the left column. I won't read through all of them, but you can see that there are a number of safeguard recommendations that came out of that memo – nine, to be precise – on the right, doing things like vetting registry operators, requiring DNSSEC deployment in new gTLDs, prohibiting wildcarding, and removal of orphan glue records – some more technical, some more procedural or process-related. So all of these safeguards that many of us are now familiar with came out this memo in 2009. Next slide, please.

Fast forward to 2016. Last year you may have seen the New gTLD Program safeguards against DNS abuse. This was written as a research aid to the Competition, Choice, and Trust Review Team. It essentially was focused on how the review team could measure the effectiveness of these safeguards. Some are easier to measure than others, but you can see the base research model that I've put on the screen here. That's included in that report, so I encourage you to go back and look at it.

You can see, essentially, at a very high level, treating the DNS expansion as an explanatory variable and the response variable

being a DNS abuse rate in some form – there's a lot of potential proxy metrics you can use for these, as you can see listed on the screen – treating the safeguards to mitigate DNS abuse as kind of intervening variables. Presumably they would have some effect on the relationship between the DNS expansion and DNS abuse rate – hypothetically, of course. It remains to be tested. Next slide, please.

Here we are. This is the current status – 2016/2017. We've just had the CCT Review Team preliminary report come out. The CCT review was an Affirmation of Commitments mandated review that specified that the review team look at malicious abuse levels. They're also mandated to look at the effectiveness of safeguards put in place to mitigate issues involved in the expansion of the TLD space.

The review team asked for a comprehensive statistical baseline measure of abuse rates in new and legacy gTLDs in order to gauge safeguard effectiveness. It also serves as a proxy for trust, such that changes in an abuse rate would presumably affect changes in trust. A very high abuse rate would probably affect how people trust a TLD. The draft report you'll note does recommend ongoing DNS abuse measurement.

The study timeline you can see on the bottom left. It's quite tight. It's tied to the CCT review's timeline. So we're trying to get

a lot done in a short timeframe. The RFP was issued in August 2016. We found SIDN and TU Delft and contracted them at the end of 2016 and began research in December. We're expecting a final report in June.

So, yes, it's a big project on a tight timeframe. As you'll note, we do need more data, especially abuse data feeds. We will definitely be calling on the community to help us out with that.

So that's the background of the study and where it is now. With that, I'm going to turn it over to our researchers from TU Delft and SIDN Labs. Thank you.

MAARTEN WULLINK: Thank you, Brian. Next slide, please. SIDN, together with TU Delft, formed a consortium to perform this study. The study was initially requested by the Competition, Consumer Trust, and Consumer Choice Review Team. Next slide, please.

The goal of the study is to come up with a statistical comparison between the DNS abuse rates in new and legacy gTLDs. We'll be looking at spam, phishing, malware, and botnet command-and-control. Next to that, we'll also be looking into potential relationships with abuse drivers, such as DNSSEC or any other drivers that are identified in future review teams. Next slide, please.

The motivation for this study was the fact that ICANN, with its New gTLD Program, enabled the addition of hundreds of new gLTDs to the root. With this program came a number of safeguards intended to mitigate rates of abusive, malicious, and criminal activity in these new gLTDs. So we'll be evaluating these as well. Next slide, please.

In order to perform this study, we're using different types of data sources. For instance, we're using domain blacklists and we're having agreements with the Anti-Phishing Working Group to provide phishing URLs; StopBadware, which provides malware URLs; and the Secure Domain Foundation, which, among other things, provides malware URLs and phishing URLs. Next slide, please.

We also need WHOIS data in order to be able to map abusive domains to registrars and registrants. For that, we have been provided with WHOIS data from WHOIS XML API, which was contracted by ICANN to provide this data. This data contains every new gTLD and also a subset of all the legacy gTLDs.

We also acquired three years' worth of zone files. We have one zone file per gTLD per day for the entire three-year study period. Next slide, please.

For the purpose of this study, we divided gTLDs into two separate groups: the legacy gTLDs, which are TLDs such as .com,

.net, and .org, or the gLTDs that have been added before the New gTLD Program, and the new gTLDs that were part of the New gTLD Program. The study has been divided into several components. Depending on the data source for each component, we were not able to use the same number of gTLDs.

For instance, if you are going to do a TLD-level aggregation of abuse, we're basically just using zone file data. We have zone files for 17 legacy gTLDs, so we're able to use these. But when we are going to look at maliciously-registered versus compromised domains or do registrar aggregation, we also need the WHOIS data. The WHOIS data only has information about nine different legacy gTLDs, so there's a slight difference there. Next slide, please.

There are also some limitations regarding this data, mainly with the collection method. We're assuming that the provider is creating a snapshot of every domain in existence at a certain point in time and then starting a scanning period. Within this scanning period, newly-created domains are not included in the list of domains to be scanned. So we might be missing short-lived domains that get created and dropped again within this period. That's something we have to investigate.

This is also very important for us. Right now, we're using quite a number of abuse feeds already, but we'd still like to include

more data in these studies – data such as phishing, malware, botnet C&Cs, and spam. If anybody has this data or knows how to get this data, please come and talk to us.

Also, if you have uptime data – this is information about the lifetime of a malicious domain name; for instance, the time between its inclusion in a black list and its removal or when it's cleaned by a registrar or a hoster. So this is also critical for our study.

Once again, please, if you have this type of data, please come and talk to us after our presentation.

BRIAN AITCHISON: Can I jump in?

MAARTEN WULLINK: Yeah. Sure.

BRIAN AITCHISON: One of the difficulties we've been encountering is the historical abuse data. There's plenty of abuse data feeds out there, but not a lot that seem to aggregate this historical 2014-2016 abuse data. I don't think there's much of a financial incentive for these abuse providers to store all that data because they get asked for

it probably once or twice for studies like this. So that's really what we're looking for. Thanks.

MAARTEN WULLINK: Thank you. Right now I'd like to turn it over to my colleague, Maciej.

MACIEJ KORCZYNSKI: Thanks, Maarten. Now I will discuss a bit more the methodology that we're going to use in this project. To study the distribution of malicious content across different providers, we first proposed to study the occurrence of unique blacklisted domains. But although this is the most commonly used and the most intuitive metric, it also has its limitations. Next slide, please.

What's the limitation? That it does not really give an indication of the amount of badness associated with a single, unique blacklisted domain. For example, botnets nowadays use quite extensively domain generation algorithms. Imagine one second-level domain name registered maliciously by the botmaster. Using [DGA] algorithms, he creates a lot of subdomains, and a subset of them is actually used for rendezvous points between compromised end machines and command-and-control servers. Next slide, please.

The problem is that this repetition metric also has some limitations, namely that there are fully qualified domain names, that it does not necessarily give an indication of the maliciousness related to fully qualified domain names. You can imagine a compromised website. Malicious users quite often distribute malicious content, like binaries or config files of malware using the distinctive paths.

For that reason, we propose the third security metric, which is the number of unique URLs. This work in fact stems from our ongoing work with the Dutch National Police. What we've been doing there is analyzing URLs used to distribute child abuse material. What we noticed is that one fully qualified domain name can be used to distribute just one malicious photo. Sometimes it is used to distribute a lot more – tens or even many more malicious photos. So for that reason, we proposed this third security metric. Next slide, please.

There are quite a lot of security reports by security vendors' companies, but in quite a lot of cases, they consider players with the highest distributions of abuse the worst ones. It's not that simple because it is also a well-known fact that large players – large intermediaries such as Internet service broadband providers, hosting providers, and also registries – simply experience a much higher number of abuse websites and security incidents.

For that reason – maybe next slide, please – in fact here, size matters. The reputable security metrics should definitely account for the fact that the larger providers – as I was saying, security ISPs and hosting providers – experience larger concentrations of abuse. So for that reason, we of course take size into account in our security metrics. Next slide, please.

So size of the TLD can in fact be used as an explanatory factor for concentration of abuse domains, and it can be interpreted as the attack surface size for cyber criminals. Let me discuss this a little bit later.

How do we estimate the size of a TLD? We've studied a number of second-level domain names registered in each gTLD, and we do it based on zone files. There are different approaches, but one alternative would be to verify monthly reports by ICANN on domain activity.

The problem with this approach is that there are registrants that buy domains and do not associate, for example, NS servers with those domains. Then they are not included in zone files. But at the same time, they are not active, so they cannot be, for example, compromised. For that reason, we are convinced that the method based on the number of second-level domains based on zone files is the right one.

Of course, it has limitations. There is a large portion of domains in new gTLDs with NS records that do not resolve yet. According to previous studies, as many as 16% of NS servers of domains do not resolve. The open question there is: is this is consistent over time? And another thing: is this 16% the same number for each gLTD? Most probably, of course not.

So the solution to this limitation would be: we are actually planning active managements to determine active domains or domains in use per gTLD. Next slide, please.

To determine the number of domains per registrar, we use WHOIS data. Again, we count the number of second-level domains registered in each registrar. Here we encounter even more limitations. For example, a single entry in WHOIS data can have multiple different names. We found, for example, a registrar using 52 distinct name variations.

What do we do? We actually perform an additional entity resolution step to try to group together the different names of a single registrar. What we do is take the name of the registrar and we convert it to lowercase. We delete all the additional information between parentheses. We remove special characters and some additional information to the type of entity, like corp or llc and so on. By applying those simply heuristics, we reduced the number of distinct entities by 58%.

But of course, another limitation is missing WHOIS data. As Maarten previously mentioned, for this part of the study, we have WHOIS data only for nine gTLDs and new gTLDs. Okay. Next slide, please.

Until now, we did not really distinguish conceptually between different groups or malicious types of blacklisted domains. Maybe I will start with the three definitions.

First, what is a maliciously registered domain? It's a domain registered by a cybercriminal for a malicious purpose. What's a compromised domain? It's a domain registered most probably by a legitimate user and hacked by a cybercriminal by, for example, exploiting the content management system like a WordPress installation or its vulnerable plugin.

We also distinguished a third group, and those are third-party domains. Those are domains of legitimate services that tend to be extensively misused by cybercriminals.

This is part of our previous study. It's also very relevant to this project. There we show that those are very extensively used. Those are file sharing services, blog post services, URL shortening services, and so on.

Before, I also promised that I would come back and discuss a little bit more the size of TLDs. What does it mean? For

compromised domains, as I mentioned before, the TLD size could be interpreted as the attack surface size for cybercriminals. The more domains in a hosting provider or in a registry, the bigger chance there is of getting compromised.

On the other hand, for malicious registrations, the TLD size could serve as a proxy for the popularity of the TLD. What makes a TLD popular? For legitimate users and for malicious users, for cybercriminals, the reasons are very often quite the same. One of the main reasons is, of course, price. Next slide, please.

Why is it so important to distinguish between compromised and maliciously registered domains? Distinguishing between those two groups is critical because they require different mitigation actions by different intermediaries. If we have a maliciously registered domain, it's the registrar that plays a key role in suspending the domain. If it's a compromised website, then it's more the hosting provider that should take an action. In practice, many roles, like the role of a hosting provider, registrar, or a DNS service operator is played by the same entity.

How do we distinguish between those two groups? Our assumption is that maliciously registered domains are involved in a criminal activity within a short time after registration. Using some previous research, we came up with the threshold being equal to 25 days. What does it mean? If the time between the

domain creation and blacklisting is less than or equal to 25 days, we consider this domain as maliciously registered. If it's more than 25 days, then we assume that the domain is compromised.

Of course, this approach has quite a lot of limitations; first of all, the lack of WHOIS data. If we do not have access to WHOIS data, we don't know the creation date, and then of course we cannot determine the time between creation and blacklisting.

The second thing is that cybercriminals quite often now register domains and then they use them in malicious activity one, two, or three months later, simply to increase the reputation of the domain.

Another thing is that, with the blacklisting process itself, we are not really sure if security vendors blacklist domains straight after some malicious activity happening. Maybe they need to collect some more evidence that the domain is indeed malicious.

For those reasons, we are working on a more advanced machine learning approach. That requires more features and the ground truth data. What kind of features? Apart from the time between domain registration or creation and blacklisting, also imagine that we will analyze the URLs. If the second-level domain name or sub-domains contain names or misspelled names of some

brands or they have some special characters, then they are presumably maliciously registered.

When we take a look at the path and we see that in the path we have wp-content, that indicates that most probably the website was compromised. More specifically, it was the content management system – WordPress – that was exploited.

Of course, there are others, like we will verify the IP address of a domain name. If it's with reverse proxy, then, again, it is very probable that is a maliciously registered domain and the attacker is using reverse proxy there. Of course, there are some more features.

Also, to model this, we need the ground truth data. At TU Delft, we have some experience with labeling maliciously registered and compromised websites. But it's quite an intense process, and there can be mistakes. We would be really grateful if you could share with us some data sources, some blacklists. We are aware that, for example, Shadowserver has those, where they distinguish between compromised websites and other types of abuse that are mostly malicious registrations. Next slide, please.

In our future work, first of all, we would like to incorporate more blacklist feeds. In addition to the analysis of TLDs and registrars, we are planning to do analyses for resellers and privacy proxy

providers and verify also geographical regions of registrants and so on and so forth.

We would like also to analyze the time to leave or the uptime of domain names. What is uptime? One of the definitions could be that it's the time required for the hosting provider, for example, to clean the compromised website as soon as it gets notified. So this is the time between the notification and actually, for example, suspending or cleaning the compromised website.

This is really important because this is a very complementary metric to the three occurrence metrics that I explained before because the occurrence metrics show how proactive the providers are. It shows if they do something to prevent security incidents, to prevent malicious registrations, or if they patch their software or hardware to prevent compromised websites.

On the other hand, we have uptime metrics that answer the question of how reactive the providers they are – how fast they react once the security incident actually happens.

Last but not least, we are going to perform an inferential analysis of potential relationships with abuse drivers. We are actually spending a few sentences on that, but that's, I would say, one of the key aspects of our study.

I will maybe give you two examples. We don't want to make any simple correlations, for example, between one particular feature and if it's a driver of abuse concentrations. We are going to model and we are going to take multiple features into account and verify if they jointly explain abuse rates.

Two examples. First example: imagine that there is a registrar that offers free registrations or very, very cheap registrations, and at the same time, we do not observe an increased number of malicious registrations. Now the question is: why? Then we notice that this is just a promotion for their customers. What does it mean? It actually means that, in the regression model, we need to take, in addition to pricing, registration restrictions so that we capture the right signal there.

Another example would be that we could verify if some security practices prevent the distribution of compromised domains. But what we need to take into account, in addition to security variables, a structural variable of, let's say, hosting providers or TLDs, meaning the number of domains in the registry or in the hosting provider. Only if we take a set of potential abuse drivers can we drive some more meaningful conclusions. Next slide, please.

We are expecting the final report in early June, 2017. Thank you so much.

BRIAN AITCHISON: All right. Thanks, Maciej and Maarten. Now I'm sure there's probably some questions out there. We do have a question on remote participation. John McCormac has been waiting patiently for us to answer, so thanks, John. I'll let Jean-Baptiste read it.

JEAN-BAPTISTE DEROULEZ: The question is, "A lot of the problem websites that I see in gTLD and ccTLD web usage surveys are compromised websites with link injections that have links that are not visible to users but are visible to search engines. Is the survey planning to cover this?

MACIEJ KORCZYNSKI: I would say that we are depending on blacklists. So as soon as they appear on a blacklist, then we will definitely [come]. That would be my [inaudible].

[DREW BAGLEY]: I think we can take it to the audience now. Jim, you have a question?

JIM GALVIN: Just a brief question. I saw on the slides that you were talking about wanting to build some machine learning in trying to make

the distinction between a malicious registration versus a compromised registration. But you also said in words that I didn't see in the slides there that you chose 25 days as your marker to distinguish between malicious and compromised.

I'm just curious as to how you came up with 25 days as your starting point. And could you say a little more about how you're going to evolve that as you get to the final report? What does machine learning really mean? I assume you're obviously going to be looking at the data and applying your own thoughts. I guess you'll talk all about how you did that [or if you had got to it] in the final report. So two related questions. Thanks.

MACIEJ KORCZYNSKI: Sure, 25 days? To be perfectly honest, we used results from some previous research. This paper analyzed spam domains, and they concluded that 99% of them are blacklisted after 25 days. So we just took it for granted and just ran a very preliminary analysis to verify how far we are from – and actually, we are quite far, I would say. That's why we decided to develop some machine learning methods.

So, basically, we will come up with as many features as possible, as some examples I gave before. We'll select a machine learning algorithm. We'll have some ground truth data, meaning that the data [is] labeled, and we will be able to, based on the ground

truth data and selected features, come up with a model. Based on the model, we'll be able to label domains from blacklists.

I hope that that answers your questions.

JIM GALVIN: It does, actually. A quick follow-up occurred to me here. Just thinking out loud, will you actually consider creating a value on a per-TLD basis to do some analysis? Different TLDs obviously have different behaviors that go on in them for a variety of different reasons. We often have discussions about the cost of a domain name, and there's an assertion that cheaper domain names tend to have greater activity on the negative side than on the positive side. I would suspect that that value, in terms of how many days it takes to move from malicious to compromised, would be different in TLDs with different characteristics.

I'm wondering if you would consider looking at that or if you had thought about that at all. The norm could be – yeah, average might not be the right thing because I can imagine standard deviations being quite large.

MACIEJ KORCZYNSKI: Yeah. Definitely we will not come up with some threshold like now – some crude threshold. That's why we proposed much, much more advanced methods. So, yeah, I think –

UNIDENTIFIED MALE: [inaudible]

MACIEJ KORCZYNSKI: No. I doubt that the attackers will have different strategies for different TLDs, to be perfectly honest. On the other hand, in general, if we do not take particular TLDs into account but groups of TLDs, the great majority of domains are compromised. This is a fact. But we suspect that, for new gTLDs, that might be actually a little bit different. The ratio of maliciously registered domains might be actually higher. But that's a general comment. For a moment, I will wait also for our final report.

[DREW BAGLEY]: Question at the end. Go ahead.

UNIDENTIFIED MALE: Just to follow up on Jim's question – maybe you already answered this, but I was curious – early in the presentation on one of the slides there was a comment about this research maybe being more relevant or something for new gLTDs that

were more inclined to see malicious behavior. I was curious as to if that was just a general comment, or if there's a specific criteria or even a list of those TLDs that you had outlined as more likely to see malicious behavior.

I can point out the comment if you want to go back to the deck, if I read it correctly. I think it was one of the first slides. Let's see. Go forward one. It's the fourth point in the lower left: intrinsic potential for malicious conduct. What does that mean?

[DREW BAGLEY]: Ah, okay. This is going back quite a long time. This actually didn't really go anywhere, from what I have read of the history of this. This was a question that was posed by these cybersecurity communities back in 2009: how do we provide an enhanced control framework for TLDs with intrinsic potential for malicious conduct?

The response to that was to create a draft framework for a high security zone verification program. Those discussions didn't really resolve too much back in – I want to say the early part of this decade. So I included that to give a comprehensive overview of what was discussed in that memo, but that discussion didn't pan out, from what I recall. Maybe some others were involved in that discussion as well.

MAARTEN WULLINK: I think what they mean there is, for example, if you would have a .bank, which actually exists, which would have a high value to create a fake .bank domain in there, given that some TLDs would have a higher value to place your malicious domain in them than others, do we need extra protections around such TLDs? I think that's what it refers to.

[DREW BAGLEY]: Right. Actually, there was a presentation yesterday that there's some sort of almost self-regulating or independent efforts with those TLDs that have that "intrinsic potential" for malicious conduct to self-regulate around creating these high-security zones. But there wasn't a comprehensive program established within ICANN to do it. It's done privately and independently. So that's where that is now.

BRADLEY SILVER: Hi. Bradley Silver. I'm from Time Warner. I have a question about whether or not the study was going to be looking at the hypothesis of past research that indicates an overlap between content [theft sites] and the incidence of malware, given that recent research last year by certain groups indicated that consumers are 28 times more likely to encounter malware in

visiting structurally infringing websites than in official websites or websites that are more legitimate. I think the other statistic was that one in three of the most well-known infringing services posed a risk to exposing visitors to malware.

I can provide you with a link to that as well. I'm just wondering if that was in the purview of your research.

[DREW BAGLEY]: I will defer to our researchers to discuss specifically to their methodology, but for purposes of the CCT review, when we get this data, we are going to include a literature review of existing DNS abuse studies, and we're going to include as many factors as possible, such as that, in there, because what we're looking at, as Brian mentioned earlier, is this broader notion of consumer trust, in addition to these safeguards that were put in place in order to prevent what were seen as perceived risks of expanding the DNS. For that, that would exactly go into the consumer trust category, looking at research that's been done on those issues.

I will look into that personally, but to the extent possible with you and others who have research such as that to contribute, please contribute it right now during this public comment period because that would be incredibly helpful for us: to see as many different sources as possible so that we can include that.

As you'll see right now in the draft report, it is merely a placeholder chapter, the one on DNS abuse, because of the fact that we are waiting for this comprehensive analysis to take place. Along with that data, we're going to factor in as many other things as possible because, as Brian said when he showed on one of those initial slides that spectrum, there is going to be obviously results that show whether there's more or less abuse in new gTLDs than legacies and how it's broken down. But then there are going to be all these other factors that are going to come into play with affecting that that might go beyond whatever effect those safeguards may have had.

BRIAN AITCHISON: I also want to emphasize that this really is a baseline study. The real interesting question – Jim hinted at it – is: what explains the variation in abuse in TLDs? That requires more advanced statistical analysis. But we do need to establish this baseline first before we can carry on with that.

Whether the public comment is on the CCT preliminary report or the forthcoming report on this, we really do encourage you to submit a public comment. I can be a sentence or paragraph long, just linking us to what we should be looking at and including in the report. Thanks.

[KARL]: [Karl], Spamhaus Project. A couple of comments. It looks very good. I like where this is going. For one, I see that you're basically saying that one of the troubles of determining the badness that you have is that your access to WHOIS data is either incomplete or not good enough. I think that's a very important comment to make in your studies because you're not the only one who has that problem. I think it needs to be said more clearly. I'm trying to put it as clearly as I can right now. The more people that say this in very simple, short sentences, the better it is because this is a huge limitation that is ongoing for a lot of people in determining if something is good or if something is bad. So that's one thing.

The other thing is that I would really like it if you would – but I know it's very difficult – start tracking the going price per domain in a certain TLD along with the studies. Maybe you can limit it to a certain bunch of providers or whatever. I think you'll find some interesting connections there, if you're not doing it already.

BRIAN AITCHISON: The pricing as an explanatory variable in abuse is something that is such a widespread hypothesis within the community that it's almost taken for granted as true. But it hasn't been empirically tested in a rigorous statistical model, which is what

we are exploring doing. There are ways that we've talked about that you can finesse the pricing data and talk about pricing levels and corresponding rates of abuse, rather than singling someone out and saying, "Hey. Your priced there and you have this abuse rate," because, as Maciej hinted, everyone has promotions. Not everyone has the same abuse rate. So it gets back to that question of what explains the variation. That's definitely something we're looking at, and we get a lot of similar comments to that.

I think we should move to the remote participation. I think John has some comments and questions.

JEAN-BAPTISTE DEROULEZ: Yes. Just for information, there are two comments and two questions in the Adobe room. The first comment is from John McCormac from HosterStats.com. "The resolving figures vary across the new gTLDs. Some of this is down to Chinese accreditation for the gTLD and registrars, and there's a development curve, a delay between a domain name being registered and a website appearing on the domain. There's also a different spectrum of domain name lengths in a TLD with a domain generation algorithm being used. It's hard to see in big TLDs but there's abuse in smaller TLDs."

His second comment is, "It is possible to detect some compromises by comparing a sites' historical web graph with the current web graph. A compromised site will gain a number of links from outside the site's social network. It can be done quickly on small TLDs but runs into scalability issues when doing it in real time."

BRIAN AITCHISON: Do you want to read John's comment? I think we'll take John's and then move back to Krisztina.

JEAN-BAPTISTE DEROULEZ: Krisztina? Okay. Sure. His last comment is, "There does seem to be a connection between low domain registration fees and infringing websites and video streaming, but it seems to be geographical/market specific."

Finally, the question from Krisztina Lanki: "After checking for abuse, would it be possible that ICANN, through the UDRP process, would give a recommendation of compensation – for example, give the minimum – to prevent unwanted actions in the future?"

DREW BAGLEY: Krisztina, that's a tough one. I don't think we're prepared to answer at this stage of the study. Again, I'll probably sound like a broken record on this: submit public comments so we can keep a comprehensive record of this and perhaps consider it in future iterations of studies or separate efforts.

Eleeza?

ELEEZA AGOPIAN: This is Eleeza Agopian with ICANN. I think that might be a question you may want to raise with the Review of All Rights Protection Mechanisms PDP Working Group, which is considering UDRP in addition to others.

TIMOTHY CHEN: Hi. Timothy Chen with DomainTools. Since you asked for this in the public forum, our company has an extraordinary amount of WHOIS data. We'd be happy to contribute it to this study if asked. Drew knows how to reach me. We have a long history of pro bono support for research like this.

I'd also encourage you – I don't know if you are, but obviously Spamhaus is here. They have an extraordinary amount of fantastic data that I think has been well-vetted. Hopefully you're reaching out to some other folks in the industry because I think you'll find a lot of support. Thank you.

BRIAN AITCHISON: I did hear you say pro bono in the public forum. No, we really appreciate that. I'll ask our Spamhaus representative: do you store this historical data?

[KARL]: Yes. The engineer says yes.

BRIAN AITCHISON: I have business cards, so let's talk afterwards.

KAL FEHER: Over time, you'll have events such as a DGA might be reverse-engineered or a software vulnerability might cause a bunch of innocent infections. How do you propose to acknowledge those historical events over the period of data capture so that you don't accidentally result in a high incidence of one or the other, simply because we may have become more technologically successful in detecting it at a particular moment in time?

Let's take an example. If today we reverse-engineered a DGA, suddenly the number of maliciously registered domains would spike. But those may well have been registered over time, or the incidents may not have increased. It's just that our ability to detect them has increased. How do you propose to not

compromise your baseline data? The same would also apply for a software infection, for example.

MACIEJ KORCZYNSKI: We simply make an observation over a longer time. That's what I could answer here. Plus, we compare, for example, two groups: the legacy gTLDs and the new gTLDs.

I will give you an example from the Anti-Phishing Working Group data. There are spikes because of those reasons that you mentioned in a very particular case last year because of an increased number of the usage of URL shorteners. Then you see the spike in both groups. So we do not really take also absolute or relative scores but also the difference between those two groups. We will definitely take this into account in [inaudible].

KAL FEHER: What I assume you're saying is that you assume that, over a long a period of time, our capabilities remain in proportion to that of the abuses, and that any increase is therefore –

MACIEJ KORCZYNSKI: But also I have some counter-examples because in 2014, I believe – you maybe remember we proposed three occurrence security metrics, and the one of them is the number of fully

qualified domain names. Just one phishing complaint that involved 13 or 17 second-level domain names in just one gTLD resulted in 32,000 fully qualified domain names.

Of course, there might be biases, but then we account for them by statistical analysis plus also a manual analysis so somehow I hope that responds to your question.

MAARTEN WULLINK: Excellent question. I'll add two things to it. Let's say a domain generation algorithm creates 50 domains a day. The miscreants will only register or two. They won't register all 50. So you'll have lot of domains that won't resolve. If some security researchers decide to step in, they are relatively easily found and they'll use different registrars and different WHOIS data than the actual bad guys will. So they're fairly easily distinguished between ones that fit within one DGA that are the original bad guys' ones and the others that are done by security researchers. So from a statistics point of view, I don't think this will be a huge issue.

DREW BAGLEY: Let me approach this DGA thing from a slightly different direction. You talk about uptimes as part of your data feeds. I guess you're referring to the Anti-Phising semi-annual report. They obviously just did a recent one here, where the goal here is

driving down uptimes of malicious activity. So you're going to be incorporating all of that into your study, I assume, and the work that you're doing.

I think that that covers the DGA activity if you properly account for all of that. The DGA activity – yeah, some of them get registered, but ultimately, if you discover what's going on, then they're not getting registered anymore. So the hit is there, but I think that the fact that you would reserve and set aside – that's what registries do; they simply set aside thousands or tens of thousands of names so that they can't be used. Now, you can't see that in your statistics, but as he's suggesting over here – I apologize; I didn't catch your name – you can certainly track down and find out about DGA activities and those algorithms and law enforcement, and you can take note of those events.

It certainly is typically public knowledge that a bunch of names were reserved. If you're creating a timeline of things here, you can insert in your timelines that these events occurred at these places at these times. You can also find out what registries participated because generally this is fairly public knowledge. You'll get some references to how many names were protected.

A lot of this makes it hard if you're trying to create algorithms to algorithmically measure things, but you can certainly make the data visible to people so that it might not factor into your

waiting schemes, but at least somebody who's trying to evaluate the data has that to look at.

I apologize. I kind of rambled a little bit. What did I want to say? The DGA stuff. I think you should go look at the DGA stuff, and you should document that. I observed that the DGA activities, I would think, should affect the uptime calculations that you're doing because people who get ahead of these things will have fewer names that are abused and visible. You should take note of the fact that there are registries who are part of these things. I think that will be helpful to the metric the guy in the back was looking for. It's certainly not a perfect thing, but it ought to at least make some data visible for people.

MAARTEN WULLINK: I'll also add that there's some excellent research being done in Germany in Fraunhofer that might be helpful for you.

BRIAN AITCHISON: We're actually scheduled until 12:15. I think everyone is excited for lunch. But, yes, we have another question or comment, please?

[ELISA]: [Elisa], INTERPOL. I noticed that in one of your slides that you mentioned that the current analysis didn't cover the privacy and proxy service. You said "given that the data is available." What do you mean by "the data is available"? Are you saying that the current P&P service registration is not labeled in the WHOIS database?

DREW BAGLEY: It has been a bit of an issue confronting us on how to parse out privacy and proxy services from WHOIS data. It's a bit difficult to categorize it, if I remember our conversations about two months ago on it. We're still trying to work out a method for determining how we can categorize it and analyze it within the privacy proxy field, but I don't know if there's more to say. It's a difficult problem that we don't quite have a solution yet.

MACIEJ KORCZYNSKI: Yeah. I would again point to the limitation of the WHOIS data, simply. We are doing our best to extract as much information from the WHOIS data as we can, but there are certain limitations.

[ELISA]: In the future, this P&P service could be covered. That would be very good. I would recommend to do so. Thank you.

MACIEJ KORCZYNSKI: Thank you.

BRIAN AITCHISON: Any more questions from the audience or in remote participation?

It doesn't look like it. I think we might be able to wrap up early and go have that lunch.

Again, I just want to emphasize that we really are calling on the community to help us find this historical data to integrate into the study. We very much appreciate public comments and links that can be very short just so we can catalogue what we should be including in our literature review, in our research. So please do participate in that. This is a big study. It's an important study. We're happy to take your input and feedback on it.

I have business cards. Drew is from the review team. He's happy to talk to you. You can find Maciej or Maarten or really – Laureen just left – the review team, too. They're happy to talk with you about this. So expect a report, hopefully a preliminary report, within a couple months, depending on how much data we can get, and a final report soon after.

Thanks very much, everyone.

**[END OF TRANSCRIPTION]**