ICANN85 Mumbai | CF – Joint Session: ALAC and OCTO
Wednesday, March 11, 2026 – 13:15 to 14:30 IST

SUSIE JOHNSON

My name is Susie Johnson, and I am the Remote Participation Manager for this session. Please note that this session is being recorded and is governed by the ICANN Expected Standards of the Behavior and ICANN Community Anti-Harassment Policy.

During this session, questions or comments will only be read aloud if submitted within the Q&A pod. Interpretation for this session will include English, French, and Spanish. If you would like to speak during this session, please raise your hand in Zoom.

When called upon, virtual participants will be given permission to unmute in Zoom. On-site participants will use a physical microphone to speak. Please state your name for the record, the language you will speak, if speaking a language other than English, and speak at a reasonable pace. I will now hand the floor over to Natalie Tercova, ALAC member.

NATALIE TERCOVA

Thank you very much, Susie. Welcome and thank you all for joining this session organized by the At-Large community together with ICANN's Office of the Chief Technology Officer, OCTO. My name is Natalia Terkova and I will be moderating today's session. This session actually continues the dialogue that began during our joint

meeting at ICANN 83 in Prague, where the At-Large advisory committee and the broader ICANN At-Large community had the opportunity to engage with OCTO on technical issues related to DNS security and abuse mitigation.

For us as the At-Large community, protecting Internet end-users and their interests is a central priority. And one of the areas we continue to follow closely is how the Internet ecosystem can improve responses to DNS Abuse, particularly through proactive approaches that help prevent harm from malicious domain name registrations before the damage occurs.

So, with this, today's session is intended to deepen our understanding of the work being carried by OCTO's Security, Stability and Resilience team. And it also builds on the discussion from our capacity building webinar that you may have attended. It was held on the 26th of February. In case you would be interested in going back, there is a recording and you can find a link on the Wiki page.

Our goal today is to continue the conversation and explore the technical perspectives behind DNS Abuse mitigation. And with this, it's my great pleasure to welcome Sion Lloyd, Principal Security, Stability and Resilience Scientist, and also Sam Cheadle, the Machine Learning Engineer at ICANN, who are joining us online. We will start the session with an introduction to OCTO, its mission and operations. Sion, over to you.

SION LLOYD

Thank you. Could we have the slide deck up, please? Perfect. And if we go to the next slide. Hello. Yeah, so I've already been introduced. My name's Sion Lloyd. I work in a team called the SSR Research Team. And along with my colleague, Sam Cheadle, we're going to talk about two of the pieces of work that we do.

I'll be talking about how we look at the underlying data that we use to build a lot of our statistics and tools upon. And Sam will be talking about another use of that data, which is associated domains, but he will introduce that for himself.

Before we get into that, I'm just going to briefly introduce the team and tell you a little bit about the way we approach projects and the way we approach problems, how our mindset works and what influences we have with our thought processes.

Next slide, please. Yeah. So, this is the introduction then to the team. Anyone who was at the webinar on the 26th of February, you'll have seen these slides before, but obviously nothing's changed in that two weeks. So, this is the department of the Office of the Chief Technology Officer, and you can see four broad teams there.

We have a research team, we have technical engagement, we have the operations side, and then we have us, the SSR Research Team, so Security, Stability, Resiliency, Research. And you can see we're a relatively small group; you're going to meet half of us today. We

all come from an academic background, so we all bring that kind of scientific process to any problems that we look to solve.

So, when we think about DNS security, DNS insecurity, DNS Abuse, we think about the way that there's lots of component parts, there's lots of different parties and players involved, and quite often it's those interfaces between those parts where the security issues can arise. And so, there's lots of places, there's lots of avenues, vectors, where there are opportunities for malicious actors to take some kind of action that is detrimental to the end user or to the internet as a whole.

And so, when we think about those actions, we think about the creation of phishing sites to steal credentials, login credentials, for maybe banks or online shopping sites. We think about the delivery of malware, so pieces of software that do something on your computer that isn't something that you want it to be doing, maybe stealing passwords again. Or things like command and control of botnet networks, so sending out messages that force a botnet to maybe launch a denial-of-service attack, for example. So, these are the sorts of things that we're looking at when we think about DNS Abuse.

Next slide, please. So, we also think at a broader level as to why these forms of abuse exist. And we have a couple of quotes here, and these are not specific to DNS or even technologically based abuses. These are from a paper on economics. And the idea here

is that security failure is caused when incentives to take action are misaligned.

So, the theory is that it's as often caused by those incentives as by poor design or maybe coding flaws. There's a more fundamental level where these things are becoming problematic. And the other quote we have here is that systems are particularly prone to fail when the person guarding them is not the person who suffers when they fail.

So, that's an example, if you like, about these misaligned incentives, where the incentive to take protective action is misaligned with the harm caused by whatever action is happening.

Next slide, please. So, one of the things that guides the way we think and guides the work we do and the projects we take on is how can we better align those incentives? And the way we see it is that metrics are a good way to show what is happening, but also to hopefully improve the alignment between action and incentive.

There's important things that we have to think about when we talk about metrics, however, because if you measure the wrong thing or if you even measure the right thing, but in the wrong way, or even if you measure the correct data, but you display it or you communicate it poorly, those can actually do as much harm as good.

So, what we're looking to do in our team is develop better metrics, develop metrics that properly measure and display the things that

we're trying to, and also then to communicate those metrics in ways where they can be understood and interpreted correctly.

Next slide, please. So, this is kind of an overview of the way we think about data and then metrics. So, data we think of as the raw material that we're working from. So, that could be DNS lookups. It could be the zone files that we have access to. It could be the reputation Block Lists that I'm going to talk about in a minute. All of that can be useful on its own, but the way you get real value from that data is to use it in different ways.

So, maybe combining two datasets, maybe running algorithms to remove entries that you're not interested in, maybe, you know, doing something that takes that data and means it makes it more-easy to communicate. So, visualizing it in different ways, for example.

All of these things are places where we think as a small team, we can actually make a difference and improve things. Okay, next slide, please. So, that is the introduction to us as a team. I'm going to go straight in then to talk about how we look at reputation data, and I'll explain what we mean by that. And then at the end of that section, I'll take questions, and then we'll have Sam talking about his piece of work on associated domains. Hopefully that works for everyone.

Okay, next slide, please. So, we have a lot of terminology, and some of it we use almost without thinking. So, quite often, and I may do this through this presentation, so I apologize. We talk about

Reputation Block Lists or RBLs. I'm going to take a little step back from that and be a little bit more generic and talk about reputation data in general. So, this is data that describes behaviors. So, it could be the behavior, in our case at least, this could be the behavior of a domain. A fully qualified host, of a URL, an IP address, and so on.

And quite often when we look at this stuff, it will describe certain behaviors like this domain has been seen to be involved with a phishing campaign, or this URL has been seen to be distributing malware, for example. All the combinations you can think of. Reputation data can also be positive, though. It could be this domain has been seen to be very popular recently.

This domain has been seen to be in the top 100 of all domains that we've seen, for example. So, it doesn't have to be negative reputation, although that is the majority of data that we deal with.

So, the initial use of this data was largely to block spam delivery. So, you download a list of domains that were involved in sending spam, and you would tell your mail server just not to accept anything that came from one of those domains, thereby protecting your users. It has evolved. There're far more use cases than just that. Now people use it to protect networks from malware, to warn about phishing, for example, if you see those pages that sometimes come up saying the page you're about to go to may be involved with phishing. You can maybe click through the link if you really want to see it.

All of the data behind those processes we would call reputation data or reputation Block Lists. And there's lots of different providers of this data, and they all have a different focus on what they're looking at. So, some may be more interested in spam, whereas others are looking at phishing specifically or just malware, or maybe they look at everything. They will have different collection methods. So, maybe some focus on mobile delivery, so SMS delivery of phishing. Maybe some only look at email delivery. So, they'll all have different strengths and consequently different weaknesses.

And we also see different models of delivery. So, when we think of reputation Block Lists, we think of a service where you can download an entire list of maybe domains, maybe URLs. And the provider is telling you that all of these are currently suspicious or have been seen to be delivering spam, for example.

Some modes are more like a point in time report. So, you don't get a complete list of what the provider believes to be bad at that point in time. You get a combination of a timestamp and a domain, and you get told that at this timestamp, there was a report on this domain that it was involved with phishing.

And so, you have to think about what it is you want to do with the data and which of those two is more useful for your particular use case. And we also see different licenses. So, some are commercial. You have to pay in order to get full access to the data. And some are open source. And maybe there's a hybrid model where for free,

you can get a certain level of access, but you have to pay to get full access, for example. So, there's lots of different ways these things can come about and can be delivered.

We also have different use cases for this data. So, for example, if you want to protect your network and block anything that's bad, or maybe you just want to put together reports on how much bad stuff has been seen recently. Maybe if you're reporting, you're interested in broad statistics. So, stuff that's aggregated at a very generic, high level. Or maybe you're actually more interested in focusing down on a specific domain or a specific malware family or something that drills through the data and looks at just a smaller portion of it.

Lots of other different ways that you might be using the data. But the point is that all of these different uses bring with them different requirements, different requirements on how you're going to use the data, what you need the data to look like in order to perform the function that you're asking of it. They'll bring different issues with them.

So, different problems. And the biggest problems we see are what happens when things don't go the way we expect. So, what are the modes of failure for our data? Is it more important that we get the full list, or are we, if we're only interested in a small number of domains, maybe we don't care if half the list is missing or a large portion of the list is irrelevant to our uses? And what are the costs of those failures?

So, one thing we think a lot about is false positives on the list. So, that would be a report of a domain that is not correct. For whatever reason, a domain is reported as, for example, phishing, but actually it's not. There's no phishing on that site. Maybe it's already been cleaned up. Maybe the phishing was never there in the first place.

So, that's what we would call a false positive. And of course, the reverse of that is false negatives. So, that is items that should be on the list, but are not. So, no list is comprehensive. No list has every single phishing domain, every single malware delivery domain.

So, what is the cost of all those missing entries? What will that impact? How will that impact the system that you're using this data for?

Next slide, please. So, the point we get to is that there is no kind of one size fits all where we look at this data. All the different providers, all those different mechanisms, all those different focuses, they all have different strengths. And because we're not collecting this data ourselves, we don't control how it's done. We don't control the methodology. We don't control any post processing that the provider does to remove entries that maybe shouldn't be there.

And the other thing that we particularly deal with, because our use case is quite, I won't say unique, but it's niche. It's a much smaller use case than the providers actually envisage. It's not what most of their customers use the data for. We have to manage how we

then deal with the data to make it suitable for our use case, because the provider won't do that for us. There's no reason that they should. They provide data to their customers. What the customers do with the data is up to them. So, we have to be prepared to do our own processing after we get the data.

Next slide, please. So, we think a lot about which data sources we should be using. Even if there's no financial cost to getting hold of the data, there's a cost on our time to develop the processes, to ingest the data, to normalize it into our system so that it can be loaded into a single source that we can then extract it from.

So, even free data has a cost. So, we have to think quite carefully before we add new sources. What data do we need? So, maybe we have a particular weakness or we believe we have a particular weakness in malware data, for example. So, we want something that will add and fill in that gap. Maybe we think we have a particular weakness geographically. So we're not getting very much data from, say, South America. So, we look for a source that concentrates on that continent so that we can fill that gap.

Next slide, please. We think about how much data will a particular source add. So, you think about that kind of cost-to-benefit part. You know, is the amount of time and financial outlay actually going to be justified when we get the data into our system? All of these things form part of the method that we use when we're looking at choosing a new data source to add to our current collection.

So, to add to that, when we think about adding a new data source to an existing collection, we have to think, well, how much does it overlap with what we already have? So, this is a fairly complex graph. But what it shows us is that some sources, they must be reading other sources underneath and then re-reporting that. So, we see a lot of Open-Source feeds, for example, being read in by commercial sources. They then do their own processing to make sure that they're happy with all those entries.

They meet their own standards of reporting. And then they will include them. And that's absolutely fine. They'll normally tell you that they're doing that. But it's important for us to know because maybe it influences our decision then on whether it's worth actually reading those underlying sources or maybe we trust the provider who aggregates many sources to have done that job for us.

It also is something that's important for us to understand. If we think that two reports, so a report on the same domain, but in two different sources, they may not quite be as independent as we believe. So, does it mean that two people have independently seen the behavior that's being reported? Or is it a re-report of some underlying source?

We also look at time-based measurements. So, if we have two sources that have the same domain or the same URL appearing, which one has seen it first? Because obviously the faster that thing is reported, the sooner we can get into our statistics or you can get

it onto a block list or do something about it. And your protection will be that much earlier.

And we also think about how much a source adds and removes new entries. Because removing old entries as things get cleaned up, as the harm is removed or as the domain is suspended, for example, removing those entries is quite important to us as well because we're reporting statistics and we don't really want to have older, no longer relevant entries in the data that we're looking at.

And sort of linked to that, how long do entries remain listed? Are they periodically rechecked by the source to make sure that they're still involved with whatever behaviors are being reported? Or do they just get put on a list and left there for 28 days and then removed?

These sorts of processes that we don't control, they're quite important to the way that we can use that data and the way we think about that data. We have some kind of less tangible processes or information about the sources, things that we can't maybe plot in a chart or measure with a simple number.

So sometimes we have to take a small sample of a list of an RBL of a source and do something to maybe look for false positives. We can't look through the whole list, we can't scale, but maybe we can look at a thousand entries and just get an idea of, are there any entries that concern us? Are there things that we wouldn't expect to see?

Also, if we know about how they collect data, do we worry about potential blind spots they may have? So, that could be they only look at SMS-based delivery or they only look in a particular geographic region. So, that may be useful. That may fill a gap where we think that we're missing coverage, but it's important for us to understand what those limitations might be.

And then the kind of more prosaic or system-based issues that we may have, is the data always available? So, if we're building a production system on this data, we need some kind of guarantee that when we ask for the data, it will be there. Is the data consistent? So, do we always get URLs or is it a mixture maybe of URLs and IP addresses? Is that something we need to understand and deal with as we read and ingest the data?

So, the final bit we think about is when we get this data, we don't have to just treat it then as right, that's the data, we're going to use it as it is. We can run our own processes on that data. So, for example, we can remove entries which no longer resolve. So, we have gTLD zone files. If the domain that we're looking at isn't in the zone file, then it won't resolve and we can remove it from statistics.

We can also look to remove things like URL shorteners, hosting services, things like Blogspot, for example, where the harm is maybe hosted on a very small portion of that domain. Maybe it's only a single subdomain underneath a larger service. And we wouldn't want to tar the whole service because there's one subdomain.

So, there may be cases where you want to deal with that stuff. But if we're looking at statistics, for example, we probably wouldn't consider those domains. We can remove domains that are suspended, but not through their removal from DNS. So, if we see certain name servers or certain redirects involved, we know that those domains have been suspended perhaps by the registrar.

Maybe they're asking for registrant detail confirmation while the report is being investigated. We also see some domains, they get sinkhole. So, this is where control of the domain is taken over and they're redirected to a site where data can be collected to look at maybe who has been affected by a piece of malware or looking at attribution of where these things are being problematic. So, we can remove those again from our statistics because they're not actively harmful. They've been mitigated, just not through suspension from the DNS.

Another thing that we're looking at is differentiating between malicious registrations and compromised domains. So, where we see, let's take phishing as an example on a domain, is there evidence that this domain was created purely for the purpose of that phishing, or is it a legitimate domain that maybe has a flaw in its content management system, which a bad actor has abused to inject their own content onto that domain? So, the way that we would deal with that, those two different cases are very different.

So, I'm just going to sum up by kind of saying a few about general observations, particularly around use cases of data. If you want to

use reputation data, if you want to use these Block Lists, think about what it is you're going to do with the data. And specifically, the most important thing is the cost of failure mode.

So, what is your cost of a false positive being on that list versus the cost of a false negative? Because that will influence whether you filter down to like the high confidence entries maybe, or perhaps it doesn't matter that there are entries on the list that aren't actually malicious. It's fine. You're not going to pay a large price if those entries are on there.

One thing we've seen is there is no one source that covers everything. Basically, every new source we add gives us many, many thousands of new unique entries. So, if you need good coverage, be prepared to combine multiple sources and think about how those different sources complement one another.

And finally, don't stop testing. Don't look at all of these things when you first start using something. And then just assume that it's going to stay that way because things change. Providers change their methodology. They change the way they deal with the data. They add new sources underneath. You know, maybe they add a whole new set of spam traps in a different geography. And don't get this data thinking that's it. I can't do anything else with it because you can always improve the data based on your specific use case.

So, with that, I'll invite questions on this section before we move on with Sam's part on associated domains. Thank you.

NATALIE TERCOVA: Thank you very much, Sion, for your presentation and also for these key takeaway points. This is Natalia Tercheva again. And let me check the room in case there are any questions. And I'm also monitoring the Q&A pod along with Yeşim in case we will have any questions submitted there. Are there any questions? Yes, Ahitha, please go ahead.

IHITA GANGAVARAPU: Hi, everybody. I am Ihita Gangavarapu for the record. First of all, thank you so much for the presentation. I think it was, I truly enjoyed all the points that you mentioned, including the coverage of sources and the various validation techniques that could be followed.

So, you've mentioned about various sources. I think in the chart that you showed, there was the anti-phishing working group. There was also Spam House, among others. And you mentioned techniques like sampling, consistency checks, and removing of the stale entries used to maintain the data quality.

So, I just wanted to understand when we look at collaboration or support from particularly threat intelligence providers in the private sector, how do you see them contributing additional signals or data sets to strengthen monitoring of abusive domains? And

from your perspective, what do you think should be the validation frameworks that could help in the integration or like a direct integration of the feeds, threat intel feeds for this particular data monitoring? Thank you.

SION LLOYD

Yes, thank you for the question. So, this is a really good example of one of the mismatches we see between the way we use the data and the way the provider imagines the data to be used. So, let's take an example of protecting a network against malware.

So, you have a list of domains that have been seen to deliver malware. And you download that from the provider. Imagine maybe 10% of those have already been mitigated. So, the harm has gone. If you're using that list to protect your network, those 10% don't matter.

You can leave them on the list. You'll never visit them if you do that you would get an NX domain response probably anyway. So, you don't have to worry about those extra entries. And the provider doesn't need to constantly check. Yes, is it still providing malware? Yes, it is. I'll leave it on the list.

And in fact, their incentive is to leave it on the list for longer. Because if they remove it, and then the domain gets reinstated and the malware comes back, they're no longer providing the protection. So, if we think about how that mismatch can be managed, it's our use case that causes that 10% to be a problem.

So, it's us who has to deal with that, not the provider. There's no incentive for the provider. And there's no requirement on the provider to remove those.

But if we can understand which 10% can be removed, then we can improve our statistics and reporting. So, I don't think it's down to the commercial sector to all the providers to make that determination. I think it's down to us. They can, of course, help. And they do help. There's a lot of sharing of methods, of sharing of all of their procedures underneath the data that they provide that they will talk to us about so we can understand it and treat it accordingly.

So, I'm not sure if that answers the full question, but it's the way that we see our use of that data specifically.

IHITA GANGAVARAPU Yeah, it does. Thank you so much.

NATALIE TERCOVA Perfect. Thank you so much. There is a question in the room by Jonathan. But I also note that we do have two questions from Aftab and Andrew in the Q&A pod. So, just please bear with me. We will go to Jonathan and then to your two questions online. Jonathan, please.

JONATHAN ZUCK

Thanks, Sion. Jonathan Zuck here. I've often wondered, we've had discussions in the DNS Abuse space for many years about the value and feasibility of predictive analytics, much as .eu attempts to do. And I wonder, one of the reasons, arguments against it are the differing economics of different registrars, for example.

But I wonder with all the data collection you're doing and the analysis you're doing, are you seeing sufficient trend data that there's some possibility that some kind of predictive analytics could come out of ICANN that could be in turn used by registrars as part of the registration process, as opposed to just being trend data as it kind of seems to be today?

SION LLOYD

Yeah, an interesting question. We see predictive data as exactly that. It is inherently different to an observation of behavior. And provided the data is clearly marked as predictive and used accordingly, we see no problem in that data forming part of the wider data sets. We wouldn't include it in reporting data.

We wouldn't include it in the statistics for domain Metrica, for example. But predictive data definitely has its place. And some of the stuff that we're looking at is predictive. In fact, you could almost think about the next presentation as a form of predictive data where we're looking at associating domains, so inferring behaviors from other observations. So, long as it's very clear that

this data is predictive, then we have no problem with it being used accordingly.

JONATHAN ZUCK: I guess my question is, are you, in your observation of data, noticing characteristics of registrations that are common among maliciously registered domains, such that the work you're doing could somehow be leveraged by those attempting to do predictive analytics?

SION LLOYD: So, I think Sam's presentation will touch on that to some extent. So, yes, we do see patterns and markers that at least indicate something is more likely to be used maliciously in the future.

NATALIE TERCOVA: All right. Thank you so much. In the interest of time, I'm now closing the queue for this part. And I will ask Yeşim if you can kindly read those two questions in the Q&A pod as they both relate to the RBL dataset. Thank you.

YEŞIM SAGLAM: Thank you, Natalie. I hope I'm audible this time.

NATALIE TERCOVA: Yes.

YEŞIM SAGLAM

No issues with the audio. Okay, perfect. So, the first question is from Aftab Siddiqui. Aftab is asking, do you provide normalized RBL dataset for general public to use? And I'm going to read the second question as well.

The second question is from Andrew Campling. Andrew is asking, hi Sion, looking at the RBL report, have you considered also evaluating its lists from commercial providers such as IncoBlox, domain tools, or DNS filter? Thank you.

SION LLOYD

Okay. So, the answer to the first question is no, we don't redistribute the data that we're using in its raw form. We only, for the majority of people, for the majority of the community, you can have access to statistics via domain Metrica. But the underlying data, because it's a combination of commercial licenses, we don't have the license coverage to redistribute that data whole.

What we can do is for registry and registrar operators, we can give them the lists of domains that are pertinent to those particular entities. But no, we can't redistribute the whole dataset as a whole. We do, we have written a document, I think it went up in the chat, which has little bits of code to do all the, some of the comparisons that we do, but that's as far as we can go.

And then Andrew, to answer your question, we look at other sources, we look at, so we re-evaluate what we have and whether

we could introduce some new source. But we don't have unlimited resources to do that. So, we have to be very careful about choosing which sources we can afford, which ones will enhance what we're doing and which ones we have the time to actually evaluate.

So, we're aware of other data that we're not currently reading, but we don't have the financial resources to load everything, unfortunately.

NATALIE TERCOVA: Thank you so much, Sion. And I believe with this, we can move to the associated domains and I can turn the floor to Sam Cheadle.

SAM CHEADLE: Thank you very much, Natalia. Good morning, afternoon, evening, everyone. My name's Sam Cheadle, the newest of the FSI team. I've been in the team for about a year and a half, working as a Machine Learning engineer.

So, I'm going to give an overview of one particular project that we've been working on over the past six months or so. This is termed very generally associated domains. So, I'll give more detail and background. There's some connection to the current PPP work that's going on. So, hopefully I can touch on that and elaborate a little bit later.

Next slide, please. So, we are looking at, oh, I'm sorry. Could you go back one slide? Just to give a little bit of the background. There's been a large amount of research on the preferences of

malicious actors and attackers around domain name abuse and domain registration.

One of the projects we were recently involved in, some of you may have come across or may have read the paper referred to as INFERMAL, carried out by KOR Labs, highlighting a number of characteristics that appear to attract malicious actors, characteristics around registrars and the services they offer.

One of those services is the ability to register domains in large volumes. And the term bulk registrations are often used. It's often debated exactly what we mean by bulk. How many domains do we need to include to use that term? I'm not going to get too much into the details there. Just to say there is evidence, there's research that the ability to use services such as APIs to programmatically register large amounts of domains quickly and efficiently is something that attracts malicious actors.

One of the questions we want to answer is, what are the characteristics of those domains? Where are they registered? So, looking across the landscape, looking across registrars and registries, can we spot where the abuse levels are highest and where these large volumes of domains are really being generated?

One of the questions that has faced us, I suppose the main obstacle we've had to tackle is, what type of data can we use to analyze this, to answer this question? The current PDP, the Working Group that's looking at the issue of associated domains is focused on how registrars can use internally held private data to find when there is

evidence, a well-evidenced abuse report, can they then pivot off that, perhaps on account holder IDs, payment records, any of that type of information, can they use that to find associated domains?

From an external perspective, from our perspective at ICANN, we have none of that information. We only have public domain registration records. So, exactly the same type of information that's available through a WHOIS or an [inaudible – 00:50:04]. So, I'm going to describe some analysis that we've carried out using public data. I'll explain a little bit more about the data source, but I'd just like to be clear, this is a very different perspective from a lot of the discussions and the valuable work that's going on around what registrars can do using their internal data. So, this is from an external perspective.

Next slide, please. So, I'll just walk through the pipeline of our data collection process and flesh out the type of algorithms that we've been running and how that feeds into this overarching goal of identifying associated domains. Sion has just described our ingestion process for a number of the different feeds, our reputation Block Lists, RBLs.

That's one of the main sources for this analysis. So, we are pulling in data, lists of domains from many different sources. We are using that in this particular analysis as a source of seeds. So, that's the jumping off point. We have a single abuse report and then we try to expand from that to see if we can find connected or associated domains.

The domain registration data I just touched upon. We're using this public domain set of registration features. No personal data, no registrant data. We're using features such as the registrar, authoritative name servers, creation dates. So, this is equivalent to what you'd see in a thin registry data set.

We are using a centrally compiled database. Each gTLD registry has an agreement with ICANN to periodically supply domain registration data. We compile that into a centrally usable source. As I mentioned, no personal level data. So, it's all that high level thin registry data.

Lastly, we are also looking at DNS resolution data. If we perform a DNS query, where does the domain resolve to? In this project, we're just looking at IPv4 resolutions. So, we want to look at all of the domains we're interested in. What is that domain to IP mapping?

The key part, or hopefully the most interesting part of this project is the next stage. So, after we've compiled that data, we apply a series of different algorithms referred to as batch registration, bulk registration, and bulk resolution. So, I'll go into some details in a second to explain those. But essentially, these are algorithms that we've designed to be open, transparent, explainable to the community, and that use layers of evidence to hopefully build up a sufficient level of evidence to reach a high enough threshold that we can make quite confident claims that we're picking out bunches of domains that are clearly associated.

Next slide, please. So, I'll just walk through a few examples that hopefully will clarify each of these different approaches or different algorithms. So the first one is the batch registration detection. So, this is a very simple illustration, all fake data. So, I've made up my own registrar. So, it's called Sam's Domains. And we're looking at a set of domains registered through that registrar, all resolving to this single combination of name servers. So, Bob at myprovider.com, Alice at myprovider.com.

And next slide, please. And what we see in this next slide is a series of domains listed along the Y-axis. They're all, as I've mentioned, all artificial data. And then we have the time of registration along the horizontal. As you can see, there are a number of unrelated. And they're just dummy coded domains background registrations that occur sequentially. So, there's no relationship, all of those red dots. And then we have a set of domains, all marked in blue that occur very close in time, almost simultaneous registrations.

The algorithm that we have focused on uses a clustering methodology to pick out high density regions of this domain registration space. So, we're just looking at the timestamps for this particular combination of domains. So, we filtered down to one registrar, one set of name servers.

Often that is sufficient then to pick out clusters of domains using an algorithm referred to as DB scan. There's a density-based clustering method. Once we've identified a group of domains that

appear to have been registered together very close in time, there's a second stage of lexical analysis.

Checking for consistency. So, in this example, we've got a number of dot top domains. They share some similarity in terms of their second order properties. So, there's no overlapping substrings in the domains, but we can see they're all six characters long and they're all a mixture of letters and digits.

As we all know, that's only one particular case that only captures one particular case, bulk registration when the domains have registered very close in time. Often, and perhaps more commonly, threat actors may spread their registrations out. It's very easy to do programmatically using an API. They may jitter the registration times. So, you won't see this very obvious, clear signal of a large group of domains registered almost simultaneously.

So, another approach that we've taken here referred to as bulk registration is to try and pick out those domains that are connected. And the way this works is to use a seed domain from one of the RBL lists that Sion described and use that as a jumping off point.

So, we have our candidate list, which consists of all the domains listed along the Y-axis here. And then we perform a clustering step where we look at all of those candidate domains, look at the lexical properties, the domain string properties, and we are picking out

related or associated domains. We can see the four related domains in this example.

And the final of those three distinct approaches is covered in this illustration. Referred to as bulk resolution. Note here that we are not looking at registration times. We are now looking at where the domain resolves, which V4 address. And instead of registration time on the X-axis, we have first observation time. And I should just describe the data set that we're using here. Primarily, these first observation times are taken shortly after the first inclusion in CT logs. So this lookup of domains and newly observed domains is happening when a domain enters a CT log.

Next slide, please. The same sort of logic is applied where we pick out a seed domain from one of the RBL lists. And exactly the same as the previous example, we're taking this candidate list, we're performing a second stage of clustering, picking out those associated domains based on this single seed domain. So, these three examples are very simplified, but hopefully they are useful just to get a general sense of what's going on.

So, I'll move on in the next part just to describe some results and some of the basic findings. Next slide, please. Just going back to that schematic, showing the pipeline of the data ingestion and the analysis, I've given a few illustrations and examples of the algorithms. I also just wanted to highlight some examples of the output.

Next slide, please. This is primarily to make it very clear the types of associated domains that these approaches are picking out are intended to be clearly and obviously related to a human observer. So, I've got a few categories. These are just three examples I picked out where we have a single seed domain for each and then I've listed the top five associated domains.

So, we've got a few different forms of association, if you like. The first one, all the associated domains are sharing at least one word. There are some features that are used in the clustering approach where we're splitting the domain string into separate English words. We're looking at the frequency of those words in terms of their use in the English language. Domain sharing rarer words are used as a very high level of evidence. They contribute to high level of evidence.

In the third column here, the first order substring, we can see these are real malicious domains registered last year. These particular ones were part of a campaign targeting UK winter fuel payments, but you can see there's a clear substring that all of these domains' share, UK-DWP. But we're not only capturing those cases where there's a clear first order overlap, you know, a shared word or a shared substring.

The methods are also capable of capturing these second order overlaps. So, that's more traditionally like what you'd see in terms of the output from a single DGA Domain Generating Algorithm. So, no kind of clear substring overlap, but clear shared properties in

terms of the length of the string, perhaps the number of letters and number of characters.

Next slide, please. Okay, so just briefly to give a use, an overview of the results. What we wanted to look at was primarily what proportion of the reported abuse domains were used or registered in bulk. And could we apply any of these techniques or do any of these techniques help in theory with proactive detection?

So, going back to one of the questions just now. Next slide, please. Top level finding perhaps most interesting or at least easiest to understand is over half of the entries that we looked at in the RBL lists. Note that we're only looking at gTLD entries in the RBLs. We're only looking at newly registered, which we presume to be maliciously registered domains, registered within six months of inclusion in the RBLs. Over half, 56.4% were tagged as associated, meaning that over half of these RBL domains could be related to at least one other domain using any of the three methods described.

So, I think that's consistent with a number of past findings highlighting the prevalence of these large-scale campaigns and bulk registrations. I won't go too much into the details of this graph, but we're looking over the course of one week. We can see that dashed black line is showing the number of seeds, so the number of domains that we get each day from the RBL feeds. And then the dashed red line is showing the proportion of those that we have tagged as associated. And you can see also in the top right there, there's a Venn diagram showing the overlap.

So, there is substantial overlap between the methods. Which I think is expected. We would be worried if there was no or very low overlap.

Okay, in terms of time, I'm moving on. Next slide, please. So, just to summarize our key findings.

Firstly, I've already described in the previous slide, over half of the RBL domains have been tagged as actual bulk. So, associated in some way. A couple of other findings that I'm just going to summarize here. If anyone's interested, this should be coming out as a report, an OCTO doc in the next few months. We looked at data acquisition delays.

So, this was one of the most interesting parts of the analysis we looked at how long it takes us to gather the data, particularly the registration data. And what are the limits that that data acquisition process imposes on proactive or real-time detection in terms of the ability for these outputs to be used by registrars' registries for suspension, for instance. We also looked at expansion analysis.

So, that's how many additional domains unreported, unknown, not listed in any of the RBLs were detected through these methods. We did find a fairly substantial volume of those. One example to pick out was high volumes of .top DGA domains. We are looking at how consistent those domains are in terms of abuse. Some of the batches were very large. For instance, we might see 10 or 20 .top

domains listed in RBL feeds, but only, I'm sorry, but a large number, perhaps hundreds or thousands, registered within a batch.

Next slide, please. I'm aware of the time, so I'm very quickly going to go over this. I just wanted to relate the findings back a little bit to the PDP, ongoing PDP that the GNSO Working Group is involved in. I think there are some interesting questions that we are interested in helping and facilitating that discussion.

Firstly, public-private data. We have this particular perspective I've outlined here. We're looking at public data only. How could that help registrars? How could it help them with their own internal checks? Could it perhaps be used as a minimum baseline? The data that we use, BRDA data, it's referred to, registration data. How are we able to output that, perform our own analysis, and give that back to relevant members of the community, relevant stakeholders?

I think I won't go into the details of the last two. The methods are designed to be flexible, not static, so we try to combat the possibility of attacker strategies changing. And just to highlight that these approaches build up a layered level of evidence. So, we are taking that approach to generate high confidence results. And specifically, those layers are infrastructure association and lexical association. Thank you. I think that's all I've got. Sorry, that's taken a little longer, Natalie.

NATALIE TERCOVA	It's okay. Thank you so much, Sam. And yes, unfortunately, we have reached the time allocated. However, I want to assure you that I have been talking with the staff, and they may email all the questions that were submitted. Or I believe, Sam and Sion, that people can also reach out to you via the email that we can see on the screen if they have any additional questions. And with this, I thank you very much, both to you, Sion and Sam, everyone in the room. Thank you to the staff and the interpreters. And I hope that we will have more joint sessions in the future. Thank you again. And this meeting is adjourned. Thank you.

[END OF TRANSCRIPTION]