Whois Misuse Study Webinar
17 December 2013 at 12:00 UTC
ICANN Transcription

Mary Wong: Minutes past the hour and so I suggest that we begin. Good morning everybody. My name is Mary Wong.

I'm a member of the ICANN policy staff. I'm here with a few of my colleagues. And I'd like to welcome you to today's Webinar on the Whois Misuse Study by Nektarios and Nicolas Christin who have been our researchers on this project for a while.

My colleague Nathalie Peregrine is the one who is running the lines. Nathalie do you have some preliminary comments for our participants?

Nathalie Peregrine: Thank you Mary. This is Nathalie. Just to remind everybody that there is audio streaming only in the Adobe Connect room.

So if you wish to ask a question over the audio during the question and answer session please don't hesitate to dial into the audio passcode is Whois.

If you have any issues during the call don't hesitate to type on private chat to me and I'll be happy to help. And back to Mary.

Mary Wong: Thank you Nathalie. And for those of you who have not been to two or Webinars before welcome for those who have been before you will know the drill usually.

So we will spend maybe the first 40 minutes with our presenters who will take us through the study, hypotheses, the methodologies they used and the findings as well as some of the analysis they've come up with.

There will be a question and answer session at the end at which point your mics will be un-muted and you can feel free to ask your questions.

If at any time during the presentation you wish for a clarification please feel free to type that in the chat as well. We will be monitoring that.

So with that welcome again. And I'd like to begin by reminding us how we got here. And you may remember that several years ago the GNSO Council resolved to obtain some objective and quantifiable data on various important aspects of the Whois system.

And as a result some studies were commissioned as well as surveys. And you see them listed here. So I will not go through all of them.

All the reports are out either for public comment or the studies have been completed already. And so this particular study on Whois misuse will complete the project that the GNSO began a few years ago on various projects on Whois.

We have a link there. And the slides will be available after the seminar today so that you can explore the studies and the work that's been done to date.

For this particular project on Whois misuse we have very fortunate to have CyLab from Carnegie Mellon University join us to perform this study.

This is some of the credentials of CyLab. I think many of you will know of their work. And of course they've done many more studies and have many more publications and research grants compared to this.

The leader that we have with us today is Dr. Christin, who is an Assistant Research Professor at CyLab and from CMU's Department of Electrical and Computer Engineering.

I've included the link here so that you can check out his bio as well as his multiple publications and awards.

And he has assisted in this project by Nektarios Leontiadis who I believe will be presenting their findings to us today.

So welcome Nicolas welcome Nektarios and over to you.

Nektarios Leontiadis: Thank you Mary. This is Nektarios. So in this Webinar we will be presenting the main methodological aspects and key results of the study conducted here at CMU in the response to ICANN's decision to pursue a set of five Whois studies as Mary mentioned characterizing the use and misuse of several aspects of the Whois service.

The empirical findings of those studies intend to inform further decisions at ICANN in terms of the future of Whois.

Now the present study tries to empirically verify by apophysis of public access to domain registrant information through Whois leads to a measurable degree of misuse.

Furthermore if this hypothesis is validated this study aims at identifying the main forms of characteristics of Whois misuse and the effectiveness of and the harvesting mechanisms in protecting registrant information.

The scope within with the study tries to answer those questions are defined by the terms of reference for Whois studies.

More specifically we look at the top five generic top level domains that in 2011 represented about 98% of the total rates or domains.

The same scope also applies to the other four Whois studies that ICANN has commissioned. Methodologically this work is divided into two main components.

The first component is a descriptive study that uses a set of interviews and surveys to capture the details of past instances of Whois misuse.

All surveys were done independently by CMU with explicit assurances of response privacy and anonymity.

The second component the experimental study tries to measure empirically and systematically the occurrence of Whois misuse. And ideally collaborate the findings were the ones from the descriptive study.

I will start with a discussion of the various components of the descriptive study and the associated key results. And then I will continue with the experimental part.

As part of the descriptive study we constructed three questionnaires targeting three key groups. The intent was to capture different but complementary perspectives on the occurrence of Whois misuse.

The first survey targeted registrars asking them in what ways if any the information listed in Whois has been misused in the past?

More importantly we certified whether they could attribute this misuse to the information being listed in the Whois.

The second survey targeted registrars and registries maintaining the zone files of the five top level domains with the intent to get aggregate data on the occurrence of Whois misuse and to identify the methods they used to protect their registrants information.

The third component was an extra survey targeting cybercrime researchers, law enforcement, and consumer in data protection organizations aiming at getting a more holistic perspective on online crime in general and more specifically on the occurrence and significance of Whois misuse.

Now starting with the (unintelligible) survey we built a representative registrant sample based on a set of 6,000 randomly selected domains from the five top level domains which we then reduced into a microcosm of 2,900 domains having the same TLD distribution as the population of domains in those five top level domains.

Given the desire for statistical significance we set the response target to 640 registrants. But despite our repeated and intense attempts to reach this goal we managed to collect 57 responses which brings the margin of error up to 12.7%.

While this is not as bad as it may seem initially as our intention is not so that Whois misuse is beyond a certain rate but rather to a certain whether the misuse occurs at a statistically significant level.

And that’s exactly what we discovered. Based on the collected reports Whois misuse occurs at the statistically significant level.

Almost 44% of the registrants participating in our survey were affected by one or more types of Whois attributed misuse.

That is misuse that the registrants could directly associate with the exposure of their personal information with the Whois.

The participants reported three main types of misuse postal address, email address, and phone number misuse.

The first two represented by postal and email spam are the most prominent ones with about 30% of the participants reportedly being affected. Phone number misuse follows with about 12% of the participants being affected.

Other types of reported harmful acts either occurred at the very small scale like blackmail or the registrant could not associate the harmful act with Whois misuse.

Now I will move on to the second survey targeting registrars and registries. The registrars we considered for participation in this survey were the 107 registrars associated with the domains in the registrant sample.

We also included the four registries responsible for the five top level domains within the scope of this study.

Overall we got 22 responses from registrars and a similar response from registry. Moreover the registrars and registries participating in the survey opted to leave many questions in the survey unanswered.

In the surveys - survey we did not make any claims of statistical significance. But the responses do represent the views of 22 of the 107 most popular registrars and one top level domain registry.

Now moving on to the findings of the survey registrars and registries reported email spam as the most prominently reported incident of Whois attributed misuse.

That is followed by phishing attempts, postal spam, malicious software sent via email, attempts for identity theft and various forms of blackmail.

Six participants reported being able to verify that they reported incidents were caused by Whois misuse.

While these results are based on the reports received at the registrars and registries it is important to know how often they directly observe of Whois harvesting and what they do about it.

Thirty percent of the participants reportedly observed attempts of Whois harvesting but no participant stated that any such attempt was successful.

Also 57% reportedly implemented one or more Whois and they have their harvesting methods. The methods implemented include IP blacklisting, grey rate limiting on Port 43 which is the designated Whois querying port, (unintelligible) and private or proxy registration services.

I need to stress out once again that the big tank of the participants did not offer any answer to these set of questions. So these findings are more indicative of the current situation rather than representative.

Before moving on to the third survey I will talk briefly about the small experiment that we used to test the registrars and registries for the existence of Whois and their harvesting techniques.

We found that about 51% the majority of registrars and registries did not employ any Port 43 rate limiting technique.

And the remaining portion uses a variety of methodologies similar to the ones reported in the survey.

Now moving on to the extra survey we build up on our contacts here at CMU and on ICANN's contacts to assemble a panel of security researchers and consumer and governmental organizations whose specialties would involve awareness of specific incidents of Whois misuse.

As I mentioned earlier the goal for this survey was to get a holistic perspective on online crime in general and more specifically on the occurrence and significance of Whois misuse.

In total we recruited 101 participants mainly security researchers and law enforcement agents. Moreover while the invited participants - we invited participants from all geographic regions we mainly have representation from the Americas and Europe.

Overall they said details on 23 incidents of Whois misuse with almost half of them targeting directly the participants.

Now most cases involved spam email with marketing material or bills but we also had a few reports or more sophisticated attacks.

In about half of the reported cases the victims did not take any protective mechanisms protective measures after the attack while the other half reportedly attempted to avoid further misuse by deploying specific countermeasures like IP blocking.

In terms of the portion of the Whois information being misused the email address has reportedly the highest occurrence with 17% of the cases followed by the registrar name and postal address in 26% of the cases.

Finally the phone number was misused in 17% of the cases.

And now we are going to move on to the experimental part of the study. And I will start by discussing the overall methodology.

The goal of the experimental study was to measure empirically and systematically the occurrence of Whois misuse. To this end we registered 400 domains across the five top level domains using a representative sample of registrars.

Each of those domains was associated with an artificial red zone identity. And the domains were crafted in such a way as to be categorized in one of five categories of interest.

Over a period of six months we used this experimental platform to measure the occurrence of Whois misuse as it was reflected by the number of email, spam, postal mail spam, and voice span collected as voicemail.

In selecting the registrars would use to register the experimental domains we started with the 107 participants sorry registrars associated with the domains in the registrant sample.

And we picked 16 based on a set of criteria we set. A key criterion was the popularity of the registrar as it was reflected through the representative registrant sample.

In total we registered 25 domains per registrar which essentially means five domains per each of the five top level domains.

Each of those domains was selected from one of the five domain categories. These categories are the following strings of completely random letters and numbers, strings representing personal names in the form of first name thus last name synthetic names composed by concatenating two random words from the English vocabulary, names representing businesses in ten professional categories often abused through phishing by online criminals. And finally names from four professional categories that are not known to be targeted.

Now moving on to the artificial registrant and identities as our audience would probably know a registrant identity is composed by the registrant's full name or organization, phone number, and postal and email addresses so each one of the 400 registrant identities had all those spaces for information.

We constructed the registrant names by piecing together random combinations of first and last names.

Also its identity had one public email address instead of private email addresses. The public one was published only through Whois. And it was in the form of contact at domain name .TLD. This set of private email addresses was not listed anywhere.

In terms of postal addresses we reused across all identities the addresses of three PO Box's located in the US.

Our initial intention was to use a set of geographically diverse postal addresses different for each registrant identity.

But we encountered a number of difficulties in this effort. And you may find more details on that in the report.

In regards to phone numbers we acquired 80 Skype numbers for the duration of the experiment. And all incoming calls were sent directly to voicemail.

I should mention that each phone number was reused between the experimental domains registered at the same register and under the same top level domain.

Now having discussed the experimental methodology I'm going to move on to the key findings of this experiment. And I will start with the case of Whois attributed email misuse.

Ninety five percent of all the emails we collected at the public email addresses was classified as spam and was directed to 71% of the experimental domains.

I need to stress that our methodology allows us to say with high certainty that all of this measured misuse is in tribute into Whois.

Now moving on to the occurrence of Whois attributed phone number misuse we collected in total -- and I apologize for the overlap here in the figure -- we collected in total 674 voicemails and 39 of those were classified with high certainty as Whois attributed spam.

The Skype accounts receiving those calls were associated with 30% of the registered domains experimental domains primarily under the .biz, .info, and .com top level domains followed by.net and.org.

It's not worthy that we faced a few challenges in classifying the voicemails. For example it's not uncommon that any one of us may get an unsolicited phone call only to realize that in fact the other party called our number by accident.

This means that we could not rely on any automated classification system but we've had to manually classify each and every piece of voicemail.

And now the third and final part of misuse were identified targeted the registrant's postal addresses. In total we collected four pieces of Whois attributed postal spam and 34 pieces of postal spam that was not associated with Whois misuse.

While this extent of postal address misuse does not allow for any meaningful statistical analysis the mere occurrence of Whois attributed postal spam suggests that registrants' postal addresses are in fact targeted and misused.

Now correlating the findings of the experimental study were the ones we reported in the descriptive study. We find that through both routes we identified the same major types of Whois misuse.

These are email address, postal address, and phone number misuse. In comparing the rates of misuse only in the case of the phone number misuse there is much between the measured and the reported rates.

In the case of postal address misuse we measured a much lower rate. And we believe that this may be caused by the limitation I mentioned earlier.

Finally the lower frequency of reported email address misuse can be ascribed to the difficulties the registrants may have in classifying email spam as Whois attributed.

As most of us receive significant loads the spam email and the modern spam filters do a good job in keeping it out of our site.

We further performed a statistical analysis to identify significant correlations between the measured misuse and the characteristics of the experimental domains.

We considered the following possibly contributed characteristics the existence of Whois in the harvesting measures, the top level domain, the price we pay to get each of the experimental domains, the category of the domain name, and the registers we used.

In terms of the observed phone number misused - misuse we found that the only factor that had a statistically significant contribution was that of the top level domain.

We found that .biz and .info domains were subject to more Whois attributed misuse while .org domains were subject to less misuse.

Considering now the observed email address misuse attributed to Whois we found a number of statistically significant correlations.

First the lack of Whois in the harvesting measures is linked to 2.3 times more misuse. Furthermore looking at the top level domain we found that .biz domains are again subject to more misuse while .com, .net and .org domains see fewer Whois attributed spam emails.

Moreover higher priced domains are correlated with less misuse. And the same applies for domain names denoting a person name.

And as you may see in this table we did not find any evidence to support a correlation between any type of Whois misuse and the registrars.

Now before moving to Q&A I will briefly summarize the findings of this study. So we found statistical evidence that through a combination of a descriptive and an experimental study that public access to Whois leads to immeasurable degree of misuse.

About 44% of the registrars participating in our survey have directly experienced Whois misuse which most prominently affects their email addresses, postal addresses and phone numbers.

Finally the evidence attests that Whois and their harvesting is capable of reducing Whois attributed spam email which is a type of misuse with the highest frequency.

And with that I have concluded the presentation of the key findings of this Whois study. And I'll be happy to answer your questions.

Mary Wong: Thank you very much Nektarios for that very comprehensive overview of the work that was done in the two studies or the two aspects of the studies.

For everyone here are some links to the public comment forum as well as where you can download the draft study report.

And would encourage you to take a look at the full report to get a fuller sense of what the research team did.

At this point Nathalie can wait un-mute the mics for everybody who dialed in? And as Nektarios said please feel free to either type a question into the chat, raise your hand in the Adobe Connect room, or speak away on the phone bridge.

I see that Jim Prendergast is typing in the Adobe Connect. So Jim if you have a question. I'll read out Jim's question for those who are on the phone but not in the Adobe Connect chat room. And his question is how the results of this will be used in conjunction with the EWG work in this space.

I can take a first initial very preliminary crack at that and I might ask Lisa Phifer or who is also working with us on the EWG to supplement what I might have to say.

At the moment Jim it's not something that's clear or been decided because the study results have just been published.

But as you may know and as others may know this is work that is being closely monitored by the GNSO Council.

And at the GNSO level there is a new working group that's just been formed to deal with a somewhat related but not entirely identical aspect of Whois.

And the reason I raised that that's on privacy and proxy accreditation and services is that that particular working group is working with the EWG to align some of the goals and some of the findings that have been sought by that EWG.

So my expectations speaking from the staff perspective is that the GNSO and the council will certainly consider these findings very, very seriously.

And coordinate with the EWG to the extent that they are either overlaps or complementary aspects that either warrant further study or that you could use in redeveloping and refining Whois into the new data directory service.

I'm sure that's not his detail that you are expecting but Lisa do you have anything to add to that?

Lisa Phifer: Mary I think you captured it. I guess the one thing that I would add is that Expert Working Group is trying to really cast a wide net and obtain as much data as we can from as many sources as we can this just being one of the sources.

The place where I would expect the results of this study might apply would be in looking at the data elements that are being proposed and the level of risk associated with each data element based on both results of this study as well as results of other studies.

Mary Wong: Thank you Lisa. And I noted that there are a couple of similar questions including from (Rudy) regarding the translation or transliteration of contact data.

And again that is a project is underway in the GNSO. And so (Rudy) I suppose I would give the same response and assure everyone that the council and the GNSO community will be following up and monitoring as appropriate.

There are two questions that I think are specific to the study that I will ask Nicolas and Nektarios to take if they can. The first is from Avri.

And the question is will the follow-up study - well actually we can take the first question the follow-up study. And Avri hopefully the earlier response that Lisa and I gave answers that to some extent.

But the specific question that you had are first what can we make of the disparity of the report in terms of experimental postal abuse?

And secondly how can one test for other aspects such as harassment or blackmail? Nicolas and Nektarios do you have responses for Avri on those questions?

Nektarios Leontiadis: Sure. In terms of studying harassment or blackmail the design of this experimental study involved the creation of artificial registrant identities which were not existed before.

So we cannot see whether the - these identities were used in ways that we were not expecting. And also we were also concerned with the observed rates of misuse through the major paths like email address, postal address and so on.

And therefore it’s not easy to monitor other types of use of these identities. And I believe Nicolas wants to say something here.

Nicolas Christin: Yes. And let me just complement this by saying if you want to measure something like blackmail or harassment these are activities that typically take place over a very long period of time.

And it’s I would say it’s very difficult to do this over a study this is only six months or so. For this you would need a very long a very longitudinal study over several years.

I think that the other question was about the disparity in experimental postal abuse. As was mentioned during the talk the results that we got between the survey and the empirical study we do believe that the empirical study here may have been a bit of a lower bound because we were operationally limited to using a handful of pure boxes in the US.

The reason for this is that there are very strict identity checks in most countries. And you can open you cannot open a pure box very easily without an actual proof of identity. And so this limited us a little bit in the experimental study.

Mary Wong: Thank you Nicolas and Nektarios. And Avri I noticed your thanks as well. There is another question from (Marcus).

And the question relates to I suppose an impression that the basis for the statistical analysis may seem quite small.

And I question specifically is whether the number of people in organizations that responded is sufficiently large to draw any conclusion? And so Nektarios, Nicolas would you like to respond to that?

Nektarios Leontiadis: Sure yes. In terms of the registrant survey while we did not get the expected turnaround we merely adjusted the error rate of the findings.

So it - they the sample was representative of the population of the registrants but we don’t have the desired precision that we intended to have initially.

Now in terms of the other surveys as I said during the presentation they don’t represent statistically significant findings but they rather give us an idea of the status quo. And they're more indicated or more suggestive than representative.

Mary Wong: Thank you and (Marcus) if you have a follow-up please feel free to type that as well.

I see that Avri you had a couple of follow-ups. And the first was a question. And it related to whether any content was actually put on a Web site?

I believe you mean during or as part of the study so that there would be something to blackmail the registrants about.

I believe that that would not have been within the scope of this study because we were looking really at the mentions of the data. Nektarios do you have a response to that?

Nektarios Leontiadis: That is correct. We did not put any content in - we do not associate any content with the registrant domains. And we were just looking at Whois attributed issues. And you're exactly right Mary.

Mary Wong: Thank you very much. We still have a little bit of time so please feel free to afraid to ask questions either as follow-ups to other questions that we were asked or that relate to questions that have yet to be asked?

And why take a moment to think and perhaps to type out go back to (Rudy)'s comment earlier on in the chat which were a follow-up I believe of an earlier question as to the translation in transliteration of contact data.

Not specifically relating to the use of the report in the GNSO Working Group on the issue but specifically (Rudy) you had a follow-up question as to whether transliteration of Whois of contact data that is if the Whois contact data was translated into a different language or transliterated into a different script other than the Latin script or English as the case may be would that increase the misuse?

That wasn't studied as part of this particular experiment or survey but Nektarios, Nicolas I wonder you had a comment about that?

Nektarios Leontiadis: We did not look at this aspect. We - the (unintelligible) identities that we used or created were all using Latin characters. So we don't have any observation in this in question to these questions. But it's an interesting follow-up.

Mary Wong: Thank you. So hopefully (Rudy) that answered your question as well?

And while you're thinking about possible final questions to ask of our researchers and presenters before signing up for today I will just remind everybody that the recording the transcript of this particular Webinar as well as the slide will be posted on the GNSO Web site following the second Webinar today which will take place at 19:00 UTC.

And thank you (Rudy) for the follow-up, thank you (Marcus), and Avri, and Jim for your comments and follow-up as well.

It seems that there are no further questions at this time. And so once again we have a - the same presentation Webinar will be conducted later today at 19:00 UTC. And we will post the recordings the transcript as well as the slides as soon as we can.

Thank you all for signing up and for joining in. We hope that you have found this presentation to be helpful.

As we've noted here there our public comment forums that are open. There's one specific to this particular study.

The initial public comment period will end in - at the end of December. And they'll be a reply comment period that goes through 18 January.

So once again we encourage you all and your groups and communities to take a look at the report. And to please submit any feedback, comments, and suggestions during the public comment forum through the link you see here on the first bullet point.

They will be extremely welcome and they will be very useful to the GNSO as it considers follow-up steps somewhat related to the questions that were asked today.

Thank you all very much. Thank you Nektarios, thank you Nicolas. And we can end the recording at this point. Thank you.

Nektarios Leontiadis: Thank you all.

Woman: Thank you.

END