A Study of Whois Privacy and Proxy Service Abuse

Final Report

7 March 2014

The National Physical Laboratory is operated on behalf of the National Measurement Office
by NPL Management Limited, a wholly owned subsidiary of Serco Group plc

NPL Management Ltd is a company registered in England and Wales No. 2937881 Registered Office: Serco House, 16 Bartley Wood Business Park, Hook, Hampshire, United Kingdom, RG27 9UY

Points of contact

Correspondence relating to this report should be addressed to:

David Hindley

Phone +44 20 8943 6325
Email david.hindley@npl.co.uk

Tony Mansfield

Phone +44 20 8943 7029
E-mail tony.mansfield@npl.co.uk

National Physical Laboratory

Hampton Road
Teddington
Middlesex
TW11 0LW
United Kingdom

Acknowledgements

The primary author of this report was Dr Richard Clayton of the University of Cambridge who has been collaborating with NPL under EPSRC Grant EP/H018298/1, "Internet Security".

The work would not have been possible without the invaluable assistance of Isaac Bright, Nicolas Christin, David Hindley, Fred Langford, Nektarios Leontiadis, Tony Mansfield, Tyler Moore, Rod Rasmussen and Sarah Smith. Data was very kindly made available by aa419.org, the Antiphishing Working Group (APWG), the Internet Watch Foundation (IWF), StopBadware, SURBL and some other organisations who choose not to be named.

The phone survey was carried out for us by IID (Internet Identity Inc.).

Contents

1. Executive summary

This study is one of series that seek to establish reliable evidence for various beliefs that are held about the operation of the Internet domain name 'Whois' system, which provides the public with information about the registrants of domain names.

1.1 Aims of the study

This particular study was originally proposed by ICANN in 2010, one of several that were to examine the impact of domain registrants using privacy services (where the name of a domain registrant is published, but contact details are kept private) and proxy services (where even the domain licensee's name is not made available on the public database). The exact definitions of privacy and proxy services that we used are set out in Section 5.

The initial intention was to test the hypothesis:

"A significant percentage of the domain names used to conduct illegal or harmful Internet activities are registered via privacy or proxy services to obscure the perpetrator's identity".

In April 2012 a contract to perform the study was awarded to the National Physical Laboratory (NPL), one of the United Kingdom's leading science and research facilities. The technical lead on the project was Dr Richard Clayton of the University of Cambridge.

We broadened the study because it was implicit that a "significant percentage" would be one that is measured – with high statistical confidence – to be substantially greater than the equivalent percentage for entirely lawful and harmless Internet activities. Hence we also sought to examine the related hypothesis:

"The percentage of domain names used to conduct illegal or harmful Internet activities that are registered via privacy or proxy services is significantly greater than the percentage of domain names used for lawful Internet activities that employ privacy or proxy services."

We wanted to consider what other methods might be chosen by those involved in criminal activity to obscure their identities, because in the event of changes to privacy and proxy services, it is likely that they will turn to these alternatives. Accordingly, we determined experimentally whether a significant percentage of the domain names we examined have been registered with incorrect Whois contact information – specifically whether or not we could reach the domain registrant using a phone number from the Whois information.

1.2 The types of activity we considered

The approach we took was to consider different categories of harmful activity and generate robust statistics for each category. We split the work into a number of work packages:

WP1 phishing
WP2 money laundering
WP3 unlicensed pharmacies
WP4 typosquatting
WP5 child sexual abuse image websites
WP6 lawful and harmless websites
WP7 domains appearing in email spam (SURBL domains)
WP8 domains associated with malware (StopBadware domains)
WP9 domains subject to the UDRP process

For each work package we have obtained a list of relevant URLs or hostnames for the particular type of activity and then categorised the domain names involved. The scale of

these activities differs considerably, but in every case we have collected data over a sufficiently long period to ensure that results are representative of each category and our results will have appropriate statistical significance.

Our study mainly addresses the use of domain names that have been implicated in illegal or harmful activities. The study also examines (particularly in WP6) some samples of lawful and harmless domain names to establish a point of reference, but it is important to understand that the selection we have made is not necessarily representative of the overall usage of domain names for lawful and harmless purposes.

For each set of domain names within the various work packages we collected and examined the Whois data for the domain names that were registered within the top five generic top level domains (gTLDs), i.e. .biz, .com, .info, .net and .org. The domain names in other top level domains (TLDs) were counted, but no further analysis was performed.

For the domain names where we had collected the Whois data we determined the proportion of these registrations that were using privacy or proxy services. If the domain was not using a privacy or proxy service we looked to see whether the Whois record contained a phone number for the domain registrant and if it did have a phone number we checked whether it passes some simple rules, so that we believe that it can be used to telephone the registrant.

We took a random sample of the domains which have these 'apparently valid' contact phone numbers and we attempted to ring up the domain registrants within this sample to have a short conversation with them, in their native language, to ascertain whether or not they acknowledged registering the domain.

1.3 WP1 (phishing) – the study in a nutshell

The overall results that we obtained can be seen with real clarity in the results of work package WP1 – where we examined domains that had occurred in URLs for phishing pages.

We split this work package into three, since we could analyse the URLs and determine whether the domain:

- was registered by a third party (e.g. companies set up to provide hosting services or URL shortening) and their services were used for criminal purposes;
- was registered by a legitimate business (or individual) whose website had been compromised and phishing web pages added without their knowledge or permission;
- appeared to have been maliciously registered for the purpose of phishing.

We found very striking differences between these categories when we considered the usage of privacy and proxy services and also whether we were successful in making contact with the registrant by phone or, conversely, had no hope of doing so:

	using privacy or proxy services		missing / invalid phone number		cannot contact by phone	phone contact succeeded
third party domains	13.7%	+	35.9%	=	49.6%	32.3%
compromised website domains	24.7%	+	37.0%	=	61.7%	23.7%
maliciously registered domains	31.2%	+	61.3%	=	92.5%	1.8%

The people who maliciously registered domains for phishing chose privacy and proxy services somewhat more than people who registered domains for legitimate purposes. However, when a privacy or proxy service was not chosen for a malicious registration a workable contact phone number was seldom given – and even if the number was apparently valid, we almost never managed to make contact with the registrant for our survey.

Conversely, even entirely legitimate 'third party' businesses that provide services to the law-abiding public – and occasionally for malicious purposes – use privacy and proxy services to a certain extent, and for almost half of the domains these businesses use there is no possibility of using the phone to reach the domain registrant. Of course there are many other ways of making contact with such businesses, and they would doubtless want people to use the information about contact pathways on their websites, rather than consulting Whois.

The compromised website category falls between these two extremes – these domain registrants use privacy and proxy services about a quarter of the time. Nearly two thirds of these registrants are impossible to contact by phone, and so we reached only a quarter of them for our survey.

1.4 Privacy or proxy service usage

The following table summarises the evidence we have of linkage between malicious registration of domains and the usage of privacy or proxy services. The main body of the report contains the detailed results and explains their statistical significance.

	Work package	Maliciously registered?	Usage of privacy or proxy services
WP6.4	Legal pharmacies	no	low
WP6.3	Law firms	no	low
WP1t	Phishing: third parties	no	low
WP6.6	Typosquatted domains	no	average
WP8	StopBadware domains	some	average
WP6.2	Executive search consultants	no	average
WP1c	Phishing: compromised sites	no	average
WP6.1	Banks	no	high
WP5	Child sexual abuse image websites	yes	high
WP1m	Phishing: malicious registration	yes	very high
WP9	Domains subject to UDRP	some	very high
WP7	SURBL domains	mostly	very high
WP6.5	Adult websites	no	very high
WP2	Advanced Fee Fraud	yes	extremely high
WP4	Typosquatting	yes	extremely high
WP3	Unlicensed pharmacies	yes	extremely high

The table clearly shows a correlation, in that maliciously registered domains have a higher usage of privacy and proxy services – but this correlation is not universal in that banks are above average users of these services, as are adult websites.

1.5 Reaching registrants by phone

The most useful way looking at the data we collected about the results of our phone survey is *not* to consider whether our survey calls were successful – there are several reasons for this not being a compelling analysis which we set out in the body of report, but one important

issue was that we often reached voicemail systems, or cellphones were not answering, and so we could not determine whether or not the phone number was valid.

Instead, we considered an opposing analysis – whether from the Whois information it would be impossible to reach the party using the domain name directly by phone. The impossibility would result from the use of a privacy or proxy service, from a failure to provide a phone number that can be called, or from the provision of a phone number that reaches someone other than the registrant or licensee actually using the domain.

The results of this analysis are shown in the following table. In two thirds of cases where domains were maliciously registered it is inherently impossible to use the phone to reach the registrant of the domain. There is also a wide range of likelihoods for lawful and harmless activities – but the pattern is far clearer than just considering the usage of privacy and proxy services in isolation: one way or another, those registering domain names to be used for criminal activity seldom provide valid contact information.

Work package	Privacy or proxy usage	Not possible to phone the registrant	Maliciously registered?
Legal pharmacies	8.8%	24.2%	no
Law firms	13.4%	33.6%	no
Executive search consultants	22.4%	36.7%	no
Banks	28.2%	44.6%	no
Typosquatted domains	19.2%	47.1%	no
Phishing: third parties	13.7%	49.6%	no
StopBadware domains	20.4%	51.4%	some
Adult websites	44.2%	55.1%	no
SURBL domains	44.1%	58.5%	mostly
Phishing: compromised sites	24.7%	61.7%	no
Typosquatting	48.2%	67.7%	yes
Advanced Fee Fraud	46.5%	88.9%	yes
Unlicensed pharmacies	54.8%	91.8%	yes
Phishing: malicious registration	31.2%	92.5%	yes

1.6 What we believe to be true

Our study shows that it IS TRUE that:

"A significant percentage of the domain names used to conduct illegal or harmful Internet activities are registered via privacy or proxy services to obscure the perpetrator's identity".

Our study shows that it is PARTLY TRUE that:

"The percentage of domain names used to conduct illegal or harmful Internet activities that are registered via privacy or proxy services is significantly greater than the percentage of domain names used for lawful Internet activities that employ privacy or proxy services."

More helpfully, we can say:

"When domain names are registered with the intent of conducting illegal or harmful Internet activities then a range of different methods are used to avoid providing viable contact information – with a consistent outcome no matter which method is used.

However, although many more domains registered for entirely lawful Internet activities have viable telephone contact information recorded within the Whois system, a great percentage of them do not."

2. Introduction

2.1 Who we are

This study was performed by the National Physical Laboratory (NPL), one of the United Kingdom's leading science and research facilities. It is a world-leading centre of excellence in developing and applying the most accurately available standards, science and technology. NPL occupies a unique position as the UK's National Measurement Institute and sits at the intersection between scientific discovery and real world application. The Royal Society and Royal Academy of Engineering oversee science quality at NPL. NPL is an ISO 9001 and ISO 17025 accredited organisation.

The technical lead was Dr Richard Clayton of the University of Cambridge, who has spent three years collaborating with NPL on an EPSRC grant entitled "Internet Security". Key technical input to specific work packages was provided by Professor Tyler Moore of Southern Methodist University and Dr Nicolas Christin of Carnegie Mellon University. All three of these people have published numerous academic papers on cybercrime and related topics and are considered to be experts in their fields.

Dr Tony Mansfield of NPL provided the experimental design and ensured that rigorous statistical analysis was performed. David Hindley of NPL was responsible for project management and coordination.

Datasets for particular work packages were provided by the Anti-Phishing Working Group, the Internet Watch Foundation (IWF), StopBadware and SURBL. Further phishing data was provided by companies who do not wish to be named. We are extremely grateful for the assistance of all of these organisations.

2.2 What we set out to do

The original May 2010 Whois Privacy and Proxy Abuse Study Terms of Reference,[1] as amended and approved by the GNSO Council,[2] sets out the objective of this study as being to test the hypothesis:

"A significant percentage of the domain names used to conduct illegal or harmful Internet activities are registered via privacy or proxy services to obscure the perpetrator's identity".

To be able to assist the ICANN community in their Whois policy formulation, it is implicit in this study objective that a "significant percentage" would be one that is measured – with high statistical confidence – to be substantially greater than the equivalent percentage for entirely lawful and harmless Internet activities. Hence we also consider the related hypothesis:

"The percentage of domain names used to conduct illegal or harmful Internet activities that are registered via privacy or proxy services is significantly greater than the percentage of domain names used for lawful Internet activities that employ privacy or proxy services."

Additionally, we believe that is vital to establish what other methods are currently chosen by malicious domain registrants to obscure their identities, because in the event of changes to privacy and proxy services, it is likely that these malicious registrants will turn to these alternatives. Accordingly, we have also determined experimentally whether a significant percentage of the domain names we examined have been registered with incorrect Whois contact information.

[1] http://gnso.icann.org/issues/whois/whois-proxy-abuse-study-18may10-en.pdf
[2] https://community.icann.org/display/gnsocouncilmeetings/Motions+28+April+2011

2.3 How we did it

Our approach is to consider different categories of harmful activity and generate robust statistics for each category. We have split the work into a number of work packages:

WP1 phishing
WP2 money laundering
WP3 unlicensed pharmacies
WP4 typosquatting
WP5 child sexual abuse image websites
WP6 lawful and harmless websites
WP7 domains appearing in email spam (SURBL domains)
WP8 domains associated with malware (StopBadware domains)
WP9 domains subject to the UDRP process

For each work package we have obtained a list of relevant URLs or hostnames for the particular type of activity and then categorised the domain names involved. The scale of these activities differs considerably, but in every case we have collected data over a sufficiently long period to ensure that results are representative of each category and our results will have appropriate statistical significance.

Our study mainly addresses the use of domain names that have been implicated in illegal or harmful activities. The study also examines (particularly in WP6) some samples of lawful and harmless domain names to establish a point of reference, but it is important to understand that the selection we have made is not necessarily representative of the overall usage of domain names for lawful and harmless purposes.

For each set of domain names within the various work packages we collected and examined the Whois data for the domain names that were registered within the top five generic top level domains (gTLDs), i.e. .biz, .com, .info, .net and .org. The domain names in other top level domains (TLDs) were counted, but no further analysis was performed.

For the domain names where we had collected the Whois data we determined the proportion of these registrations that were using privacy or proxy services. If the domain was not using a privacy or proxy service we looked to see whether the Whois record contained a phone number for the domain registrant and if it did have a phone number we checked whether it passes some simple rules, so that we believe that it can be used to telephone the registrant.

We took a random sample of the domains which have these 'apparently valid' contact phone numbers and we attempted to ring up the domain registrants within this sample to have a short conversation with them, in their native language, to ascertain whether or not they acknowledged registering the domain.

This study only processed publicly available personal data, gathered from publicly visible Whois data. The information we gleaned from telephone conversations with the people identified as domain registrants is only presented in statistical form.

It should be noted that this study does not attempt to do any sort of general survey regarding who uses privacy and proxy services and why this choice is made, but merely analyses the incidence of usage for domains implicated in particular activities.

It might be questioned why we chose to concentrate on phone numbers for this study, effectively using their presence and correctness as a proxy for the overall veracity of the registrant's details. Our reasoning was that we would get a significantly higher response rate

from conducting a telephone survey than if we had chosen to communicate with domain registrants using either fax or email – and that it would only be by actually communicating with the purported domain registrant that we could be sure that their contact details had not been used without their knowledge. Additionally, although grossly invalid email addresses can be detected, there are a range of failure modes that we would have missed and these would have been indistinguishable from a valid address that reached a registrant who did not wish to respond to us.

2.4 Organisation of the rest of the report

We start by considering the role of Whois in countering criminality and then we survey the results of previous studies into the incidence of privacy and proxy registrations and clarify exactly what is meant by these terms.

We then set out our methodology in detail: how we collect Whois data and how we analyse it to determine if a privacy or proxy service has been used and whether the registrant has provided an apparently valid contact phone number. We then explain the process we used for trying to make telephone contact with samples of purported registrants to determine whether they agreed that they had registered specific domains. These phone calls had numerous different outcomes and we explain how we divided these into five categories.

We then discuss each of the nine work packages in turn, explaining the source of our data and giving the results from our analysis.

We then consider our results as a whole, reviewing the differing proportions of domains in the different categories that were found to be using privacy and proxy services. We also review the wide differences we encountered in our ability to make telephone contact with domain registrants. In this discussion we also provide details about the statistical methods we used and the extent to which it is appropriate to consider differences in results to be truly significant and not just artefacts of our sampling approach.

We finish with an analysis of the results of the various work packages – setting out a broad-brush overview of what would otherwise be just a myriad of detailed information.

3. The role of Whois in countering criminality and unlawful activity

The various types of criminality and unlawful behaviour that we consider in this study are almost all directly related to the content of websites. A standard countermeasure to the creation of a criminal website or the addition of extra, criminal, pages to an existing website is to arrange for the fraudulent pages to be 'taken down'. Webpage 'take down' is achieved by communicating with someone who can suspend the web hosting and/or with someone who has sufficient access to the website to make the necessary changes.

The hosting company can often be identified by looking up IP addresses in the appropriate Regional Internet Registry (RIR) Whois system rather than the domain name Whois system which we consider here. In this case, action may result from contacting the hosting company to either deal with the problem or use internal customer records to contact the party actually using the website.

In some cases, where a website has been compromised, action is swifter if the website owner is contacted directly, and if there are no contact details on the website itself then the domain Whois information can be useful.

Some websites use so-called 'fast-flux' techniques, where the URL hostname points to a different relay machine every few minutes. For this type of attack the most practical approach is to get the hostname suspended (i.e. removed from the DNS) though this is almost invariably done by contacting the registrar for the domain name rather than attempting to locate the person that registered the domain.

Fast-flux attacks are currently rather rare and the only one we noticed during our work was a single instance of a phishing domain in the Indian (.in) top level domain.

All of this means that, in practice, the accuracy of the Whois information is only occasionally relevant when trying to counter criminality. However, if the Whois information is patently false then this can sometimes add weight to the argument that the domain name is being used criminal activity – which can expedite action being taken.

That said, Whois information can be of real importance when countering unlawful behaviour (matters dealt with under civil rather than criminal law). In particular it can be very significant to understand whether or not a particular domain has been registered by a commercial rival and there can be significant economies of scale in taking action to deal with many types of intellectual property infringement if it can be ascertained that a large number of relevant domains have been registered with essentially identical contact details. In both examples the use of privacy or proxy services makes it very difficult for a brand owner to balance the costs against the benefits when considering proceedings to defend their rights.

4. Previous ICANN studies of privacy and proxy services

In 2009 the Chicago based National Opinion Research Center (NORC) collected Whois data for an ICANN commissioned study. On 17 Jan 2010 they published a document entitled: "*Draft Report for the Study of the Accuracy of WHOIS Registrant Contact Information*".[3]

Their aim was to study a sample of domain names which were balanced as regards the country of residence of the domain registrant, but they did not wish to incur excessive costs at the stage of their project where they interviewed the registrants.

Their approach, which is described in detail in the draft report, was to collect Whois data on 2400 domains which were sampled in appropriate ratios from the .biz, .com, .info, .net and .org generic top level domains. They then examined the registrant address and extracted the country name. The countries represented in the sample were then split into stratified groups and individual countries were selected from these groups for further study.

This resulted in NORC doing a detailed analysis of just 1419 domains (from the original 2400) with Table 2 of their draft report setting out the countries involved. Of these 1419 domains, NORC eventually determined that 351 domains were using privacy or proxy services. This gives an incidence of 351/1419 (24.7%) and their sample size meant that their estimate was ± 2.2% at a 95% confidence level.

Later, on 14 Sep 2010, ICANN published "*ICANN Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the top 5 gTLDs*".[4] This second study used all of the 2400 domains from the NORC dataset, irrespective of country, and determined how many of them were using privacy or proxy services. This time the total came to 429 (17.8% ± 2.0%).

The substantial difference between these two percentage figures was not commented upon in the second report... but we understand from various conversations that NORC refined their classification methodology between the first and second studies, which led them to reclassify some of the domains.

NORC have recently (23 May 2013) published the results of a new investigation: "*Whois Registrant Identification Study Project Summary Report*".[5]

For this new publication NORC selected a completely new set of 1600 domains, once again randomly selected in appropriate proportions across the top five gTLDs and found that 320 were registered using privacy or proxy services (i.e. 20.0% ± 1.6%).

The results from the latest NORC investigation are statistically equivalent to the previous (Sept 2010) report and NORC conclude "there is no evidence to suggest that the usage of privacy and proxy services has changed over time".

[3] http://www.icann.org/en/resources/compliance/reports/whois-accuracy-study-17jan10-en.pdf
[4] http://www.icann.org/en/compliance/reports/privacy-proxy-registration-services-study-14sep10-en.pdf
[5] http://gnso.icann.org/en/issues/whois/registrant-identification-summary-23may13-en.pdf

5. Identifying privacy and proxy services

The definitions we use in this report are fully consistent with the NORC studies:

> A *privacy registration service* offers alternative Whois contact details (street address and phone number) that belong to the service. The identity of the registered name holder is not hidden, but it is not possible to attempt direct contact with the registrant using just the information within the Whois record.
>
> A *proxy service* registers a domain name in its own name and then licenses the use of the domain to a third party (a customer of the proxy service). The Whois system lists the proxy provider as the domain registrant. The licensee's identity and contact information are not published in Whois.

Both privacy and proxy services sometimes provide a registrant contact email address that is specific to the particular domain, but the registrant's usual email address is not revealed.

NORC's documented approach (set out in detail in both of the 2010 reports) was to look for organisation names that specifically identified themselves to be privacy or proxy services. They also looked at all registrant names containing words such as "privacy", "proxy", "shield", "domain" etc. This does correctly identify many privacy and proxy services, but false positives can occur, for example when organisations use role names such as "Domain Owner" for registration purposes or with India's convention of spelling out that companies are "Private Limited" organisations.

In this study, which involved the inspection of tens of thousands of Whois records so we had considerable practice, we found it was straightforward to identify the majority of cases where privacy and proxy services had been used – and in many cases the Whois information contained specific comment text that clearly indicated the use of a privacy or proxy service. However, for completeness, we also ran an additional check for the presence of any of NORC's list of keywords anywhere within the registrant details.

NORC's other heuristic was to suspect the use of a privacy or proxy service where multiple domains had the same registrant name, organisation or address. This was less effective for our study because NORC was selecting domains completely at random, whereas in several of the work packages the domain names we studied had been registered by the same people – that is they were inherently linked to each other.

Nevertheless, it still proved to be useful to manually inspect where multiple domain names had the same contact phone number. In the vast majority of cases this was clearly just someone who had registered several domains – but it did help us identify of handful of further privacy and proxy services.

Where doubt remained – mainly where lawyers and web developers registered domains on behalf of their clients – we always assumed that the registrant was not a privacy or proxy service. This choice was made specifically to ensure that any resultant bias in our results would undercount the incidence of privacy and proxy services.

6. Obtaining and analysing Whois data

Once it was determined that a domain name was to be included in our study (with that determination being specific to each work package) then it was processed as follows:

6.1 Fetching the raw Whois data

The top level domain (TLD) was identified and a count made of how many domains were registered within each TLD. This study was specifically chartered to examine registrations under the five largest generic top level domains (gTLDs) i.e. .biz, .com, .info, .net or .org. and accordingly, Whois data was only analysed for domains falling in these gTLDs.

For the work packages which used near real-time ('live') event feeds to identify relevant domains, the Whois data was fetched by a batch system that ran every 30 minutes. In some cases, Whois data was not immediately available from the registrar. If this was automatically identified (some registrars have rate limiting systems) then the domain name was added in to the next batch – so that the data was collected 30 minutes later.

For other work packages – where fixed lists of information were processed – there was just one large batch, containing all the relevant domains. These batches were processed promptly when the batch of data was received, albeit it sometimes took a day or so to collect all the data when registrar rate limiting was being applied.

When multiple failures occurred, and for the cases where the automated system failed to identify that a failure has occurred, the Whois data was obtained from a commercial Whois data recording service. This service archives copies of the Whois data as it changes over time, so it is possible to obtain a copy of what the Whois data would have looked like had it been successfully fetched at our initial attempt.

In a very small handful of cases, the domain was de-registered before Whois data could be obtained and the commercial service was not of assistance either. These cases form a separate category in the results of each work package, and we have no practical choice but to exclude them from all further analysis, and we do not count them in any of the summary statistics that we present.

6.2 Processing the raw Whois data

The raw Whois data was then passed through 'deft-whois' a Whois data extraction system developed specially for this study. The system uses knowledge of the way that each individual registrar formats their Whois responses to identify the information recorded about the domain name registrant.

The study aims to identify whether a privacy or proxy service was used when the domain was registered. The deft-whois system identifies this directly because it has been programmed to recognize the contact details used by the privacy and proxy services.

The study also aims to determine whether the registrant has provided an 'apparently valid' phone number. Some registrars provide a phone number in the Whois results along with the other registrant details, in which case this number was used.

When the registrant details did not include a phone number then the Administrative, Technical, Billing and Zone details were inspected (in that order). If there was a phone number in such a section where the other details are clearly those of the registrant – for example the street address matched – then this was taken to be the 'apparently valid' phone number for the registrant.

Phone numbers with fewer than 6 digits or where the digits (apart from a country code) were all zeroes or all nines were entirely ignored – the assumption being that the registrant had entered these values to assuage the validation functions of a web form rather than because this was actually a valid phone number for the registrant.

Other values with distinctive patterns (such as 987654321) were recorded as apparently valid (however unlikely this might be) because it would be impossible to distinguish between memorable numbers being used by businesses and numbers that were just made up.

The phone numbers were converted to a proper international dialling format – and tidied up accordingly. Whois data often contains invalid formats, where the international code is repeated in the number itself, or zeros included along with the international code. It is also common to find +1 being prefixed to non-US numbers (which we identified by considering the postal address details for the registrant).

> For example, we saw the Indonesian number +62.819xxx expressed as +1.620819xxx, +62.62819xxx and +1.0819xxx on different domains registered to the same postal address in Indonesia.

Phone numbers that were clearly invalid for the particular country because of their length or area codes were excluded, however some countries, such as Germany, do not have a fixed number of digits for phone numbers so this could not always be done. Invalid area codes (particularly the US area codes of 555 and 111) were also excluded. When in doubt, we deemed the numbers to be 'apparently valid' – although, as can be seen from our results, many of the numbers turned out not to be contactable.

6.3 Selecting which registrant phone numbers to call

A random selection was made from the domains where we had identified an apparently valid phone number for the registrant.

A sub-contractor (IID, Internet Identity Inc.) was employed to call this subset of numbers with the aim of having a short conversation with the registrant in their native language. The instruction sheet we provided for this task can be found in Appendix A.

This study is intended to determine whether the phone number recorded in the Whois results can be relied upon as a way of reaching the person who registered the domain. We expected that some phone numbers would not work, or would reach people whose details corresponded with the Whois information but who denied having registered the domain. The sole intent of the conversation with the (purported) registrant was to establish the reliability of the phone number.

We did not make it a requirement that our sub-contractor speak to the registrant in person; if the person who was actually reached was able to link the registrant and the domain then that was sufficient to show that the phone number would be a viable way of making contact.

In several cases, multiple domains had the same contact phone number. This was used, as discussed above, as one way of identifying privacy and proxy services, but these duplicate phone numbers sometimes came about just because someone had registered more than one of the domains that we studied.

We have made the assumption that the result of a phone call to the particular number would be independent of which relevant domain was called about, but that making multiple phone calls to the same person might lead to them behaving differently during the later calls. For

that reason, we excluded numbers that had been called before (whether in the same work package or not) from further rounds of random selection.

That said, in several work packages the same contact details were provided for multiple domains (sometime for several hundred domains at once). Our method of making random selections from the entire set of domains means that it is more likely that a registrant for many domains will be selected than a registrant of a single domain – but they can be selected at most once, their response to our survey about the particular domain that caused them to be chosen is taken to indicate how they would answer if quizzed about any other domain that they registered.

6.4 Scheduling of telephone calls to registrants

Having made a selection of the phone numbers to call, we used a manual process to assess whether domain registrants were likely to be a business or an individual – because we hoped thereby to improve the chances of reaching them by phone.

For businesses we prescribed the call schedule:

#1 ring during the morning of a business day

If number fails to connect at all, record this and make no further attempts. If no answer, then make up to 3 more attempts:

#2 ring during the afternoon of a business day (preferably the same one)

#3 ring during the morning of a different business day

#4 ring during the afternoon of a different business day

For individuals, or where we were unsure what sort of registrant was involved, we prescribed the schedule:

#1 ring during the late afternoon of a business day

If number fails to connect at all, record this and make no further attempts. If no answer, then make up to 3 more attempts:

#2 ring during the early evening of another business day

#3 ring during the morning of yet another business day

#4 ring during the mid-evening of a business day

6.5 Categorising the results of telephone calls to registrants

We provided a list of domains and telephone numbers as a batch to our sub-contractor and asked them to make the required calls, scheduling them as needed for whether we had indicated that they were likely to be a business or an individual. We also provided a copy of the Whois results for the domain in case it would be useful on the call.

We indicated which work package had caused us to include the domain in the study, but we did not tell them if we made any categorisation of the domain, for example, whether we had come to the opinion that it had been maliciously registered.

Our sub-contractor passed the results of the calls to us and we include these results in our statistics for each of the work packages.

For clarity, we have reported call results under five headings, concentrating on the outcome rather than the way the outcome was achieved:

1. Invalid phone number / does not connect

   This category includes all the numbers which, when called, resulted in silence or a number unobtainable tone from the telephone system. In some cases numbers were too short and in other cases they must have been entirely bogus. Note that many such cases were identified during our initial analysis, as described above, so this category consists solely of the cases where the number was 'apparently valid', but was not valid in practice.

2. Number is not answered

   This category covers numbers where a ringing tone was heard, but the call was not picked up. Note that when this happened further call attempts were made on the predetermined schedule set out above and cases are put into this category only when all of the call attempts get the same result.

3. Inconclusive call / answering machine

   This covers all the calls where someone picked up the phone but it was not possible to ascertain whether the registrant of the domain was there or not. This included the cases where the registrant was never available and no-one could speak on their behalf and cases where the phone line was too bad for the conversation to be understood. The cases where the call was answered by an automated machine were also put in this category, as were the handful of cases where the correct person was reached but they refused to discuss whether they had registered the domain or not.

4. Number does not work to reach registrant

   This covers the cases where there was a conversation with the person who answered the phone but they denied that the domain registrant could be reached at that number and they could not suggest any method of reaching the person; or there was a conversation with the named person, or someone speaking authoritatively on their behalf, but they denied having registered the domain.

5. Number works to reach registrant

   This covers cases when someone answered the phone and they agreed that they had registered the domain, or they agreed that someone else at that number had registered the domain, or that the domain had been registered for their company.

6.6 Inferring and scaling results

As already noted, we took steps to ensure we never called the same phone number twice because we have assumed that we would get the same result if we had done so. Hence for each work package, after presenting the raw results we obtained from our phone calls, we also present these 'inferred' values from the results of the calls we did make (in that particular work package or, occasionally, another one).

We then "scale up", by appropriate multiplication, from our sample of phone calls to all of the domains with phone numbers. This simplifies comparison between different categories of result (between data determined for every domain and that which was just sampled).

7. WP1 'Phishing'

Phishing is the creation of fake websites for the purpose of stealing security credentials. It is generally associated with spoof bank websites, but phishing sites now attack many other businesses such as webmail providers, online games, auction sites, ecommerce companies and social networks. The URLs of these websites are mainly circulated by email and instant messaging systems.

7.1 Raw data for this work package

A number of entities provide lists ('feeds') of phishing URLs and for this work package we have used five feeds that have been made available, for research purposes, to Richard Clayton at the University of Cambridge:

1. APWG – the Anti-Phishing Working Group (APWG) is a pan-industry body which collates the URLs submitted to them by members and by members of the public who use their web-based reporting system.

2. Phishtank – the website phishtank.com is a community site which accepts reports of phishing websites from visitors. It also provides crowd-sourced validation by having visitors vote on whether or not URLs lead to phishing sites.

3. Takedown company #1 – this feed is provided by one of the companies which sells phishing website 'take down' services. Their feed contains the URLs of the companies for which they act, as well as other URLs that they have learnt about.

4. Takedown company #2 – this feed is provided by another company which sells phishing website 'take down' services. Their feed contains the URLs of the companies for which they act, as well as other URLs that they have learnt about.

5. Brand owner – this feed is provided by the owner of several major Internet brands and contains only URLs for phishing websites that attack these brands.

For this study we considered all the new URLs received in a one week period 18 – 24 April 2012 from all of these feeds.

The raw count of URLs received was 32068, but there were many duplicates (some of the feeds incorporate a subset of the URLs found in other feeds), and there were also a number of false positives because some of the feeds fail to validate URLs prior to distribution. The duplicates were discarded and the false positives ignored.

A typical example of a false positive that had to be removed were the URLs placed into marketing emails that allow measurement of recipients' clicks on items of interest by linking to an intermediate, tracking, website. Automated systems often pick out these URLs as suspicious (the email mentions a bank, but the URL does not point to the bank website). Clayton's feed processing system contains numerous heuristics for spotting these false positives and they were augmented, for this study, by manual inspection of any URLs that did not appear in multiple phishing URL feeds.

A further type of false positive was the presence of malware URLs in some of the feeds which Clayton's processing also removed. During the sampling period a number of emails were being sent out that mentioning financial institutions, but anyone unwise enough to click on the links in those emails would have visited a website that attempted to exploit browser

security flaws – the URLs are of course malicious, but they are not 'phishing' so they were excluded from this work package. Similar URLs were of course studied in WP8.

7.2 Report inflation

A specific issue with analysing the feeds in this particular study was the 'report inflation' practiced by some URL reporters.

Some websites are configured such that the URL `http://example1.com/~a/b.html` references the same page as `http://example2.com/~a/b.html` (for all of the domain names `example1.com`, `example2.com`, etc. that resolve to the same IP address).

If the sender of the phishing emails was aware of this configuration, then they might use a wide range of URLs in their attacks – and hence the reporters feel justified in reporting not just the URL that they actually observed being used, but all of the other URLs that the phishing campaign might possibly use in the future.

In practice, the senders of phishing emails seldom if ever exploit the possibility of using alternative names. If we were to include all of these alternative names in the study this would change what we were measuring from "domains that are involved in phishing attacks" to "domains hosted on the same server as domains involved in phishing attacks". We did not feel that this would be useful and so we chose to exclude the inflated reports.

If we had been able to know which of the alternative domain names provided for the same website was the domain name involved in the phishing attacks, then we would have included only that original domain name in our study. Since the original domain name was not available to us, we excluded all URLs cited by such inflated reports

7.3 Categorising the data

The processing described above (removing duplicates, false positives and removing URLs involved in 'report inflation') yielded 16384 unique URLs which utilised 4993 distinct domain names spread across 130 top level domains (TLDs).

The most commonly used 24 TLDs were:

com	**2205**	**44.2%**
tk	375	7.5%
net	**302**	**6.0%**
br	237	4.7%
org	**195**	**3.9%**
info	**110**	**2.2%**
uk	100	2.0%
in	99	2.0%

au	98	2.0%
ru	79	1.6%
de	68	1.4%
pl	63	1.3%
cl	60	1.2%
nl	55	1.1%
mx	50	1.0%
ca	43	0.9%

ar	40	0.8%
za	39	0.8%
fr	38	0.8%
dk	37	0.7%
ro	35	0.7%
it	33	0.7%
biz	**27**	**0.5%**
us	26	0.5%

Between them, the five gTLDs being studied compromised 56.9% of the total number of domains, and accounted for 8088 (49.4%) of the URLs.

The domains being studied were then split into three groups, which we expected to show different patterns of registration.

1. Third parties (we call this subset 'WP1t' in subsequent discussion):

   These are legitimate businesses whose domain name appears in a phishing URL as a result of their services being used for criminal purposes, or a compromise of one of their legitimate customers. The businesses:

   - provide a URL shortening service (e.g. notlong.com);
   - offer free web hosting (e.g. 0fees.net);
   - provide a dynamic DNS hostname service (e.g. dyndns.org);
   - provide a cloud service (e.g. google.com);
   - or the URL uses a reverse DNS hostname (e.g. charter.com).

   The Whois data for these domains was examined in the usual way. Although one might expect that the majority of these sites will have valid contact information in their Whois, they will generally provide other mechanisms for making contact to report abuse.

2. Compromised websites (we call this subset 'WP1c' in subsequent discussion)::

   These are websites owned by legitimate businesses, organizations, or individuals, which have been compromised by the phishing attackers who then arrange for their pages to be served in addition to those of the website's owner.

   There is no *a priori* reason to suppose that these people and organizations created their domains for malicious purposes, and no reason to believe that their decisions about what information to place into Whois influenced whether or not their site was compromised.

3. Malicious registrations (we call this subset 'WP1m' in subsequent discussion):

   These domains have been specially registered for use in phishing attacks.

Identifying the domains that fell into category #1 (WP1t) was relatively straightforward. Some of the domains are household names, and there are a number of published lists of domains in the various categories. Distinguishing between categories #2 (WP1c) and #3 (WP1m) can often be very simple. It is often the case that intruders only obtain partial access to a compromised website – so that they are constrained to add their phishing pages deep within the directory hierarchy. If so then the URLs are easy to distinguish; for example, if a WordPress installation has been compromised and extra webpages added then the phishing URL might be: `http://example.com/wp-includes/images/bankpage.html`.

However, where the URL was fairly generic, a manual process had to be applied. In some cases the domain name was pretty clearly registered for phishing (for example, `statuspaypal.com`), but other domain names were rather less distinctive.

Domains where legitimate content was present (perhaps only findable in search engine caches by doing a search on the domain name) were treated as compromised, as were a number of sites where evidence was found of people boasting, perhaps months earlier, of having been able to deface the site. Where there was doubt, the assumption was made that the site was compromised rather than maliciously registered.

7.4 Results

For each category (third parties, compromised websites and maliciously registered domains) we present the results as a series of tables.

The left hand table of each pair gives the results from processing the Whois data: whether a privacy or proxy service has been used, and if not, whether an apparently valid phone number was provided for the registrant. Note that the percentages exclude the domains for which we were unable to obtain any Whois data.

The right hand tables give the results of our phone calls to a sample of the (non-privacy non-proxy service) registrants when an apparently valid phone number was present in the Whois. The first column of these tables shows the 'measured' results for the 200 calls we made from our sampling of domains in this work package (we split these 200 calls across the three categories to correspond with their prevalence). The second column shows the results we have 'inferred' by assuming that we would get the same result for all the other domains which had the same phone number as one that we called (whether in this work package or another one).

Third Parties (WP1t):

no Whois	1		
no phone		**11**	4.2%
phone		**216**	82.1%
privacy	2		
proxy	34		
privacy+proxy		**36**	13.7%
	TOTAL	**263**	

phone call results:	measured	inferred
invalid phone number/does not connect	9	57
number is not answered	9	31
inconclusive call / answering machine	1	2
number does not work to reach registrant	1	1
number works to reach registrant	20	59
TOTAL	40	150

Compromised Machines (WP1c):

no Whois	0		
no phone		**109**	5.1%
phone		**1488**	70.2%
privacy	37		
proxy	487		
privacy+proxy		**524**	24.7%
	TOTAL	**2121**	

phone call results:	measured	inferred
invalid phone number/does not connect	21	34
number is not answered	12	13
inconclusive call / answering machine	2	3
number does not work to reach registrant	1	1
number works to reach registrant	25	26
TOTAL	61	77

Malicious Registrations (WP1m):

no Whois	5		
no phone		**24**	5.3%
phone		**285**	63.5%
privacy	29		
proxy	111		
privacy+proxy		**140**	31.2%
	TOTAL	**449**	

phone call results:	measured	inferred
invalid phone number/does not connect	74	109
number is not answered	9	12
inconclusive call / answering machine	0	1
number does not work to reach registrant	12	17
number works to reach registrant	4	4
TOTAL	99	143

The next set of tables shows the results we get from scaling up our inferred results to cover the whole of each category of domains. We need to do this scaling to include the "no phone number" category so that we can correctly ascertain the percentage of calls that are likely to result in particular outcomes.

Third Parties (WP1t):

uses privacy or proxy service	36	13.7%
no phone number in Whois	11	4.2%
invalid phone number/does not connect	82	31.2%
number is not answered	45	17.0%
inconclusive call / answering machine	3	1.1%
number does not work to reach registrant	1	0.5%
number works to reach registrant	85	32.3%

Compromised Machines (WP1c):

uses privacy or proxy service	524	24.7%
no phone number in Whois	109	5.1%
invalid phone number/does not connect	657	31.0%
number is not answered	251	11.8%
inconclusive call / answering machine	58	2.7%
number does not work to reach registrant	19	0.9%
number works to reach registrant	502	23.7%

Malicious Registrations (WP1m):

uses privacy or proxy service	140	31.2%
no phone number in Whois	24	5.3%
invalid phone number/does not connect	217	48.4%
number is not answered	24	5.3%
inconclusive call / answering machine	2	0.4%
number does not work to reach registrant	34	7.5%
number works to reach registrant	8	1.8%

It will be seen that there are substantial differences between the categories as to whether a privacy or proxy service is used – and if not, whether a phone call will succeed in reaching the domain registrant.

Discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

8. WP2 'Advanced Fee Fraud and other complex scams'

The website aa419.org collates reports of websites associated with complex online frauds. Its original focus was on advanced fee frauds (often called "419 scams" after the relevant article in the Nigerian criminal code) so it contains details of fake banks and fake law firms. Additionally, it contains numerous reports of websites for fake transport and logistics companies, often associated with auction escrow scams.

Although there is occasional use of free web hosting sites, the overwhelming majority of the websites listed by aa419.org use domain names that have been specially registered by the scammers, with a name that is chosen to mislead potential victims.

8.1 Raw data for this work package

We recorded all of the 717 URLs listed by 419.org over a 28 day period (18 Aug 2012 to 14 Sep 2012), which gave us 713 domain names to study.

These domains were registered under 21 different TLDs:

com	**510**	**71.5%**
org	**64**	**9.0%**
net	**62**	**8.7%**
uk	20	2.8%
info	**13**	**1.8%**
eu	11	1.5%
us	6	0.8%

co	5	0.7%
biz	**3**	**0.4%**
de	3	0.4%
ru	3	0.4%
za	3	0.4%
tk	2	0.3%
au	1	0.1%

ca	1	0.1%
cc	1	0.1%
gp	1	0.1%
in	1	0.1%
nu	1	0.1%
pro	1	0.1%
st	1	0.1%

The five gTLDs being studied in this report cover 652 domains (91.4% of the total).

For the period we considered there was a sole example of a legitimate business whose domain name appears in one of these scam URLs as a result of a their services being used for criminal purposes (this was us.com, which allows registration of subdomains). We do not consider this domain further, but continue with the 651 other domains.

8.2 Results

We randomly selected 200 domains[6] which had apparently valid contact phone numbers for their registrants and called these numbers to determine whether or not we could contact the domain registrant. We present the results for this work package in tabular form below.

The left hand table gives the results from processing the Whois data: whether a privacy or proxy service has been used, and if not whether an apparently valid phone number was provided for the registrant.

The upper right hand table gives the results of our phone calls to a sample of the (non-privacy non-proxy service) registrants when an apparently valid phone number was present in the Whois. The first column of this table shows the 'measured' results, and the second column shows the results we have 'inferred' by assuming that we would get the same result for all the other domains which had the same phone number as one that we called (whether in this work package or another one).

[6] One domain was incorrectly categorised at domain selection time, so the actual results are for 199 calls.

The lower right hand table shows the effect of scaling up the inferred results to cover all of the domains. We need to do this scaling to include the "no phone number" category so that we can ascertain the percentage of calls that are likely to result in particular outcomes.

no Whois	0		
no phone		**10**	1.5%
phone		**338**	51.9%
privacy	21		
proxy	282		
privacy+proxy		**303**	46.5%
	TOTAL	**651**	

phone call results:	measured	inferred
invalid phone number/does not connect	126	147
number is not answered	13	18
inconclusive call / answering machine	18	24
number does not work to reach registrant	37	44
number works to reach registrant	5	10
TOTAL	199	243

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	303	46.5%
no phone number in Whois	10	1.5%
invalid phone number/does not connect	204	31.4%
number is not answered	25	3.8%
inconclusive call / answering machine	33	5.1%
number does not work to reach registrant	61	9.4%
number works to reach registrant	14	2.1%

Discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

9. WP3 'Unlicensed pharmacies'

An unlicensed pharmacy is an Internet business that sells pharmaceuticals to individuals without being licensed by any relevant body. Numerous jurisdictions deem specific lists of drugs to be controlled substances which cannot lawfully be supplied except by licensed pharmacies – often requiring a doctor's prescription to be produced at the point of sale. Unlicensed pharmacies do not make any attempt to meet these requirements.

Most of the marketing of unlicensed pharmacies is operated on an affiliate basis – whether the method of promotion might be the sending of email spam, the posting of irrelevant blog comments or any of the numerous other advertising methods employed. The spammer, blog poster, etc. receives a cut when a purchase is made as a result of their promotion efforts. The domain names that are placed into links in the advertising copy form a key part of the tracking system that ensures that affiliate payments are correctly allocated.

9.1 Raw data for this work package

The domain names we investigate come from a study by Nektarios Leontiadis and Nicolas Christin of Carnegie Mellon University. Every day from November 2011 to October 2012, following a methodology they had previously established,[7] they entered specific drug names that are commonly sold from unlicensed pharmacies into a major search engine and recorded the URLs from the first page of results. Although some of the results were for legitimate sites, the majority were links to unlicensed pharmacies – giving 832 domain names for this study.

These domain names were all registered specifically as unlicensed pharmacies with just 2 exceptions where subdomains of web hosting company domains were used. We do not consider these 2 domains further, but just analyse the other 830 domains.

These domains were registered under 21 different TLDs:

com	**633**	**76.3%**
net	**70**	**8.4%**
org	**33**	**4.0%**
biz	**26**	**3.1%**
uk	15	1.8%
ru	8	1.0%
us	8	1.0%

eu	7	0.8%
ua	6	0.7%
co	4	0.5%
in	4	0.5%
info	**3**	**0.4%**
pro	3	0.4%
at	2	0.2%

au	2	0.2%
cc	2	0.2%
ca	1	0.1%
fr	1	0.1%
it	1	0.1%
mobi	1	0.1%
ws	1	0.1%

The five gTLDs being studied in this report cover 765 domains (92.2% of the total).

9.2 Results

For 69 domains (9% of the total) we were initially unable to obtain a Whois record with details of the registrant. In almost all of these cases, a Whois record was available, but it was a generic "suspended domain" entry provided by the registrar. The high number of instances of this relative to other work packages resulted from the nature of the data collection – some of the domains had only been active for a fairly brief period anything up to a year earlier.

When we looked at the Whois data for the remaining 696 domains to identify candidates for making phone calls we found that there were only 212 phone numbers we had not already

[7] N. Leontiadis, T. Moore & N. Christin: *Measuring and Analyzing Search-Redirection Attacks in the Illicit Online Prescription Drug Trade*. In Proc. 20th USENIX Security Symposium (USENIX Security'11), San Francisco, CA. pp. 281–298, Aug 2011. https://www.usenix.org/legacy/events/sec11/tech/full_papers/Leontiadis.pdf

called in the other work packages we had processed at that point. We chose not to follow our usual methodology of making a random selection of 200 from these, but attempted calls to all 212 to determine whether or not we could contact the domain registrant.

For the 69 domains without Whois records we subsequently used a commercial service to obtain copies of the Whois which were current at the point at which they were in active use. The bulk of these 69 had been using privacy and proxy services and in a few cases we had called the number already in regard to another domain. We decided against making phone calls to the 14 new phone numbers we encountered because these domains had been inactive for many months and the delay might well have affected whether we could reach the domain registrant and how they answered our survey question.

Once again we present the results in the same tabular form.

The left hand table gives the results from processing the Whois data, whether a privacy or proxy service has been used, and if not, whether there is an apparently valid phone number for the registrant.

The upper right hand table gives the results from making a phone call to a sample of the (non-privacy non-proxy service) registrants when an apparently valid phone number was present in the Whois. The first column of this table shows the 'measured' results, and the second column shows the results we have 'inferred' by assuming that we would get the same result for all the other domains in all of our samples (from all work packages) which had the same phone number as one that we called.

The lower right hand table shows the effect of scaling up the inferred results to cover all of the domains. We need to do this scaling, to include the "no phone number" category, so that we can ascertain the percentage of calls that are likely to result in particular outcomes. Since we called all but 14 of the relevant phone numbers, the scaling does of course have only a minor impact on the values.

no Whois	0		
no phone		**1**	0.1%
phone		**345**	45.1%
privacy	11		
proxy	408		
privacy+proxy		**419**	54.8%
	TOTAL	**765**	

phone call results:	measured	inferred
invalid phone number/does not connect	160	245
number is not answered	20	39
inconclusive call / answering machine	14	16
number does not work to reach registrant	13	26
number works to reach registrant	5	5
TOTAL	212	331

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	419	54.8%
no phone number in Whois	1	0.1%
invalid phone number/does not connect	255	33.4%
number is not answered	41	5.3%
inconclusive call / answering machine	17	2.2%
number does not work to reach registrant	27	3.5%
number works to reach registrant	5	0.7%

Discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

10. WP4 'Typosquatting'

Typosquatting is the registration of small variants of the domain name of a legitimate website. This is done in the hope that a small proportion of people who intended to visit the legitimate website will miss-key the URL and thereby accidentally visit the typosquatting domain instead.

Research in 2009 by Tyler Moore and Ben Edelman identified nearly a million domains which were identical, within a typing error or two, to the most popular 3200 website names using .com (which are five or more characters long).[8] They found that 80% of the sites were hosting pay-per-click advertisements, often advertising the correctly spelled domain and its competitors.

It should be noted that typosquatting is not a crime, albeit the domain registrant might be subject to civil penalties, and that the Uniform Dispute Resolution Policy (UDRP)[9] could be invoked to settle the disputed use.

It is believed that many of these typosquatting domain names are registered by the same people. Because of the economies of scale enjoyed by a brand owner who can deal with multiple domain names in a single initiative, there is clearly some incentive for registrants to hide their identities from casual inspection and to diversify their registrations.

10.1 Raw data for this work package

For this study, Tyler Moore provided a list of candidate typosquatting domains that were current in mid-January 2013.

He consulted the Alexa global rankings table for website visits and extracted a list of the domains in .com that were at least 6 characters long (there were 3045 of these). He then looked for single typos (one character added, removed or changed) for domains between 6 and 13 characters long and double typos (two changes) for domains which were 14 or more characters long.[10] To establish whether the domain existed he consulted the .com zone file.

The raw file contained 30466 domains. Of these 313 were found to be expired, 3008 were found to have the brand owner as the registrant and 91 were identified as being the websites of legitimate third parties who just happened to have registered a similar domain name (e.g. galottery.com and walottery.com are "typos" for calottery.com, but also legitimate state lotteries in their own right).

The result of this processing meant that this work package studied the Whois data for 27053 domains – the largest group of domains that we considered.

However, it should be noted that for reasons of efficient data processing this work package was only able to consider domains within the .com top level domain and we did not consider typosquatting within the other four gTLDs that the other work packages consider.

[8] T. Moore & B. Edelmann: *Measuring the Perpetrators and Funders of Typosquatting*. 14th International Conference on Financial Cryptography and Data Security, LNCS 6052, Springer, pp.175–191, 2010. http://lyle.smu.edu/~tylerm/fc10typo.pdf

[9] The exact mechanism for resolving disputes will vary, depending upon which TLD is involved.

[10] Note that the "dot" before "com" is treated as a character when identifying typos. This is because browsers often add a ".com" to hostnames without a URL – so examplecom.com is considered a typo for example.com. Similarly, because of the prevalence of "www" in hostnames, wwwexample.com is considered a typo for example.com (and both of these domains currently host pay-per-click adverts!).

Once again we present the results in the same tabular form.

The left hand table gives the results from processing the Whois data, whether a privacy or proxy service has been used, and if not, whether there is an apparently valid phone number for the registrant. Note that the percentages exclude the domains for which we were unable to obtain any Whois data.

The upper right hand table gives the results from making a phone call to a sample of the (non-privacy non-proxy service) registrants when an apparently valid phone number was present in the Whois. The first column of this table shows the 'measured' results, and the second column shows the results we have 'inferred' by assuming that we would get the same result for all the other domains in all of our samples (from all work packages) which had the same phone number as one that we called.

The lower right hand table shows the effect of scaling up the inferred results to cover all of the domains. We need to do this scaling, to include the "no phone number" category, so that we can ascertain the percentage of calls that are likely to result in particular outcomes.

no Whois	46		
no phone		**1076**	4.0%
phone		**12911**	47.8%
privacy	378		
proxy	12642		
privacy+proxy		**13020**	48.2%
	TOTAL	**27007**	

phone call results:	measured	inferred
invalid phone number/does not connect	55	292
number is not answered	29	300
inconclusive call / answering machine	55	1015
number does not work to reach registrant	14	643
number works to reach registrant	47	639
TOTAL	200	2889

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	13020	48.2%
no phone number in Whois	1076	4.0%
invalid phone number/does not connect	1305	4.8%
number is not answered	1341	5.0%
inconclusive call / answering machine	4536	16.8%
number does not work to reach registrant	2874	10.6%
number works to reach registrant	2856	10.6%

It will be noted that there were a non-trivial number (46) domains for which we were unable to obtain Whois data. The .com gTLD employs a two-level Whois system. In all cases a valid response was received at the first level from whois.crsnic.net, but in 11 cases there was no response at all from the Whois server that whois.crsnic.net indicated should be consulted next. In the other 35 cases there was a response from the second level Whois server, but this response was to the effect that no data was available – implying that the first level response was incorrect.

Further discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

11. WP5 'Child sexual abuse image websites'

Child sexual abuse image websites (sometimes described as child pornography sites) are universally reviled and are the subject of active police investigations into the extremely serious crimes involved.

Public lists of such websites are, for obvious policy reasons, difficult to obtain. However, the Internet Watch Foundation (IWF) kindly agreed to provide data for this study. The IWF was founded in 1996 to provide a hotline service for the reporting of criminal online content. It is a founder member of INHOPE (the Association of Internet Hotlines).

In 1996 the majority of the material the IWF dealt with was carried on Usenet, but for many years almost all the material has been distributed from websites. In recent years the trend has been towards the use of generic hosting sites, "cyberlockers" and peer-to-peer distribution systems, but the IWF still encounters some websites where the domain has clearly been registered for the express purpose of hosting criminal content. Some of the domain names give an indication of the type of content that might be expected – but others are rather more anodyne.

When illegal material is encountered the IWF analysts make a contemporaneous record of the URL and the Whois data for the domain name. For this work package we considered the 656 domain names that they had encountered during the 2012 calendar year which they considered to have been registered for criminal purposes.[11]

The TLDs involved are:

com	**479**	**73.0%**
net	**69**	**10.5%**
info	**23**	**3.5%**
org	**23**	**3.5%**
in	17	2.6%
tk	14	2.1%
biz	**8**	**1.2%**
ru	8	1.2%
eu	5	0.8%
us	3	0.5%
co	2	0.3%
asia	1	0.2%
be	1	0.2%
jp	1	0.2%
pro	1	0.2%
sh	1	0.2%

For this study, just as in the other work packages, we only considered domains registered within .biz / .com/ .info / .net / .org, which were 602 in all, 91.8% of the total.

[11] Where the IWF dealt with material hosted on 'free webspace' or in 'cyberlockers' the domain name was registered by a legitimate company. We do not consider such domains in this work package.

Presenting the results in standard form gives this table:

no phone		**10**	1.7%
phone		**411**	68.8%
privacy	19		
proxy	157		
privacy+proxy		**176**	29.5%
	TOTAL	**597**	

No attempt was made to make any contact with any of the domain registrants in this category. This was because we were processing this data many months after it was current, because we wished to avoid disrupting any law enforcement activity that might be occurring, and because we do not believe that if we did reach someone who had actually registered the domain we would receive a truthful answer when we asked them about it.

However, the IWF suggested to us, drawing on their experience and that of law enforcement investigators, that if we had attempted to contact the registrants of the 411 domains where there was apparently valid contact information, we would have found that the name, address and phone number had been extracted from public records and the person had no connection with the registration of the domain.

Further discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

12. WP6 'Lawful and harmless websites'

For comparison purposes, we also wished to determine what range of variation occurs in the use of privacy and proxy services when domain names are registered for use by a number of different types of legitimate website.

The categories have been chosen to approximately mirror the criminal and harmful sites studied in some of the other work packages. However, it is important to keep in mind that these particular categories do not necessarily reflect overall usage of privacy or proxy services by the totality of all lawful and harmless websites.

We consider six specific categories within this work package. For each we describe the basis on which we have selected domain names to study and we then present the results in the usual tabular format.

In each instance the left hand table gives the results from processing the Whois data, whether a privacy or proxy service has been used, and if not, whether there is an apparently valid phone number for the registrant. Note that the percentages exclude the domains for which we were unable to obtain any Whois data.

The upper right hand table gives the results from making a phone call to a sample of the (non-privacy non-proxy service) registrants where an apparently valid phone number was present in the Whois. The first column of this table shows the 'measured' results, and the second column shows the results we have 'inferred' by assuming that we would get the same result for all the other domains in all of our samples (from all work packages) which had the same phone number as one that we called.

The lower right hand table shows the effect of scaling up the inferred results to cover all of the domains. We need to do this scaling, to include the "no phone number" category, so that we can ascertain the percentage of calls that are likely to result in particular outcomes.

Full discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

12.1 WP6.1 Banks

We wanted to consider a group of domain names that might allow further insight into the results of WP1 (phishing) and decided to consider banking websites.

We extracted all of the domain names from the "Business and Economy > Shopping and Services > Financial Services > Banking > Banks" section of the Yahoo! directory which, on 1 April 2013 gave us a list of 2019 domains. These were registered under 76 different TLDs of which the 24 most numerous were:

com	**1641**	**81.3%**
uk	32	1.6%
net	**26**	**1.3%**
au	21	1.0%
jp	16	0.8%
my	15	0.7%
ch	15	0.7%
br	12	0.6%

es	12	0.6%
nz	10	0.5%
org	**10**	**0.5%**
ca	9	0.4%
de	9	0.4%
ph	9	0.4%
hk	8	0.4%
ie	8	0.4%

in	8	0.4%
It	7	0.3%
kr	7	0.3%
lu	7	0.3%
th	7	0.3%
lv	6	0.3%
pt	6	0.3%
sg	6	0.3%

There was just one .biz and just one .info domain and hence the five gTLDs being studied in this report cover 1679 domains (83.2% of the total).

The URLs from the Yahoo! directory were then all visited to ascertain whether they were still banking websites (the directory turned out to be poorly curated and contained a number of entries for banks that no longer exist or were for organisations that were connected to the banking industry but were not banks in their own right). Where there was no website to inspect the Whois data was inspected to determine if the domain registrant was still associated with a banking organisation (this was the case for a number of defunct banks, where the successor institution had kept hold of the domain name, but was not providing an associated website).

At the end of this validation process just 1405 domains remained. From the 900 of these with apparently valid phone numbers for the registrant a random sample of 40 were selected and attempts made to make contact. The results, in standard format were:

no Whois	0		
no phone		**109**	7.8%
phone		**900**	64.1%
privacy	352		
proxy	44		
privacy+proxy		**396**	28.2%
	TOTAL	**1405**	

phone call results:	measured	inferred
invalid phone number/does not connect	3	5
number is not answered	1	6
inconclusive call / answering machine	21	31
number does not work to reach registrant	3	3
number works to reach registrant	12	14
TOTAL	40	59

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	396	28.2%
no phone number in Whois	109	7.8%
invalid phone number/does not connect	76	5.4%
number is not answered	92	6.5%
inconclusive call / answering machine	473	33.7%
number does not work to reach registrant	46	3.3%
number works to reach registrant	214	15.2%

12.2 WP6.2 Executive search consultants

We wanted to consider a group of domain names that might allow further insight into the results of WP2 (Advanced Fee Fraud) and in particular the sites that recruit "money mules", viz: who advertise jobs which involve receiving payments and forwarding money.

We initially intended to consider lists of recruitment companies, but were unable to find a substantial well-curated list with a global reach. We therefore decided to use the list of members of Association of Executive Search Consultants.[12]

When we visited this website on 11 April 2013 it listed 320 members of the Association, but this included many country-specific branches of the same multinational and there were a handful of members without a website. Thus we actually obtained a list of 257 distinct domain names.

[12] https://www.executivesearchconnect.com/eweb/StartPage.aspx

These domains were distributed across 31 different TLDs:

com	**175**	**68.1%**
br	8	3.1%
net	**8**	**3.1%**
au	6	2.3%
ca	5	1.9%
de	5	1.9%
fr	5	1.9%
se	4	1.6%
ch	3	1.2%
cl	3	1.2%
ie	3	1.2%

uk	3	1.2%
be	2	0.8%
es	2	0.8%
eu	2	0.8%
fi	2	0.8%
in	2	0.8%
no	2	0.8%
nz	2	0.8%
pl	2	0.8%
ru	2	0.8%
za	2	0.8%

ar	1	0.4%
at	1	0.4%
cn	1	0.4%
dk	1	0.4%
gr	1	0.4%
jp	1	0.4%
nl	1	0.4%
org	**1**	**0.4%**
ro	1	0.4%

We considered just the .com, net and .org gTLDs (184 domains; 71.6% of the total). However, one of the websites was using a subdomain of .ru.com, and so we excluded this 'third party' site from further consideration and just analysed the other 183 domains.

From the 132 of these domains with apparently valid phone numbers for the registrant a random sample of 40 was selected and attempts made to make contact. The overall results for WP6.2, in standard format, were:

no Whois	0		
no phone		**10**	5.5%
phone		**132**	72.1%
privacy	24		
proxy	17		
privacy+proxy		**41**	22.4%
	TOTAL	**183**	

phone call results:	measured	inferred
invalid phone number/does not connect	3	3
number is not answered	7	8
inconclusive call / answering machine	9	9
number does not work to reach registrant	2	2
number works to reach registrant	19	19
TOTAL	40	41

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	41	22.4%
no phone number in Whois	10	5.5%
invalid phone number/does not connect	10	5.3%
number is not answered	26	14.1%
inconclusive call / answering machine	29	15.8%
number does not work to reach registrant	6	3.5%
number works to reach registrant	61	33.4%

12.3 WP6.3 Law firms

We wanted to consider a further group of domain names that corresponded to the sites involved with Advanced Fee Fraud activity (WP2) – where fake law firms are sometimes used for scams involving inheritances.

We wanted a well-curated list with a global reach and chose to use the membership of Lex Mundi, which describes itself as "the world's leading network of independent law firms".[13]

[13] http://www.lexmundi.com/lexmundi/About_Lex_Mundi.asp

Member law firms are located throughout Europe, the Middle East, Africa, Asia and the Pacific, Latin America and the Caribbean and North America.

Lex Mundi has 158 members, but five do not have websites, and there were ten duplicate domain names (for example, where there were member firms in Dublin and Belfast, sharing the same domain name for their respective websites). Therefore we analysed the data for 143 domains, which were spread across 34 different TLDs:

com	**108**	**75.5%**
org	**3**	**2.1%**
ar	1	0.7%
bb	1	0.7%
br	1	0.7%
bs	1	0.7%
cl	1	0.7%
co	1	0.7%
cy	1	0.7%
eu	1	0.7%
hk	1	0.7%
hr	1	0.7%

hu	1	0.7%
is	1	0.7%
jp	1	0.7%
mx	1	0.7%
na	1	0.7%
net	**1**	**0.7%**
ni	1	0.7%
no	1	0.7%
pe	1	0.7%
pl	1	0.7%
pt	1	0.7%

py	1	0.7%
ro	1	0.7%
rs	1	0.7%
ru	1	0.7%
si	1	0.7%
sk	1	0.7%
tc	1	0.7%
tw	1	0.7%
uk	1	0.7%
uy	1	0.7%
za	1	0.7%

We considered just the .com, .net and .org domains (112 domains; 78.3% of the total). From the 85 of these domains with apparently valid phone numbers for the registrant a random sample of 40 was selected and attempts made to make contact.

The overall results for WP6.3, in standard format, were:

no Whois	0		
no phone		**12**	10.7%
phone		**85**	75.9%
privacy	14		
proxy	1		
privacy+proxy		**15**	13.4%
	TOTAL	**112**	

phone call results:	measured	inferred
invalid phone number/does not connect	5	5
number is not answered	4	4
inconclusive call / answering machine	18	18
number does not work to reach registrant	0	0
number works to reach registrant	13	13
TOTAL	40	40

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	15	13.4%
no phone number in Whois	12	10.7%
invalid phone number/does not connect	11	9.5%
number is not answered	9	7.6%
inconclusive call / answering machine	38	34.2%
number does not work to reach registrant	0	0.0%
number works to reach registrant	28	24.7%

12.4 WP6.4 Legal pharmacies

WP3 considered the domain names used by unlicensed pharmacies, so we now consider, for comparison purposes, the domain names used by some legitimate pharmacies.

On 28 March 2013 we fetched the list maintained by LegitScript of online pharmacies that they considered to be "safe for US patients".[14]

This list contained 264 pharmacies using 255 different domain names in just five TLDs:

com	**244**	**95.7%**
org	**5**	**2.0%**
net	**3**	**1.2%**
us	2	0.8%
biz	**1**	**0.4%**

We considered all but the .us domains (253 domains; 99.2% of the total) and from the 214 of these domains with apparently valid phone numbers for the registrant a random sample of 40 were selected and attempts made to make contact.[15]

The overall results for WP6.4, in standard format, were:

no Whois	2		
no phone		**15**	6.0%
phone		**214**	85.3%
privacy	16		
proxy	6		
privacy+proxy		**22**	8.8%
	TOTAL	**251**	

phone call results:	measured	inferred
invalid phone number/does not connect	6	10
number is not answered	0	0
inconclusive call / answering machine	17	65
number does not work to reach registrant	2	6
number works to reach registrant	16	31
TOTAL	41	112

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	22	8.8%
no phone number in Whois	15	6.0%
invalid phone number/does not connect	19	7.6%
number is not answered	0	0.0%
inconclusive call / answering machine	124	49.5%
number does not work to reach registrant	11	4.6%
number works to reach registrant	59	23.6%

[14] http://www.legitscript.com/pharmacies

[15] There were 41 calls made, rather than 40, because one domain was, for a time, incorrectly categorised as an unlicensed pharmacy and so a phone call was made as part of the WP3 activity.

12.5 WP6.5 Adult websites

WP5 considered the domain names used by websites that were set up to distribute child sexual abuse images, so we now consider, for comparison purposes, the domain names used by a range of websites providing 'adult' content, that is to say erotic material which would not be suitable viewing except by consenting adults.

We extracted all of the domain names from the "Business and Economy > Shopping and Services > Sex > Adult Galleries" section of the Yahoo! directory which, on 1 April 2013 contained 3758 entries, which used 3594 domains.

We excluded 14 domains because they were general purpose providers of free web space (some of these domains turned up in WP1t (phishing)), but continued to consider 2 domains which advertised themselves specifically to be providers of web space for adult material.

The 3578 domains that we analysed were registered under 34 different TLDs:

com	**3276**	**91.6%**
net	**149**	**4.2%**
org	**44**	**1.2%**
nu	13	0.4%
tv	12	0.3%
uk	11	0.3%
info	**10**	**0.3%**
us	8	0.2%
cc	6	0.2%
biz	**5**	**0.1%**
ca	5	0.1%
de	5	0.1%

to	5	0.1%
jp	3	0.1%
nl	3	0.1%
br	2	0.1%
ee	2	0.1%
ru	2	0.1%
ws	2	0.1%
ar	1	0.0%
at	1	0.0%
be	1	0.0%
ch	1	0.0%

co	1	0.0%
cx	1	0.0%
dk	1	0.0%
hm	1	0.0%
in	1	0.0%
me	1	0.0%
mobi	1	0.0%
ph	1	0.0%
tm	1	0.0%
vc	1	0.0%
vu	1	0.0%

We considered all the domains registered in .com, .net, .org, .info and .biz (3484 domains; 97.4% of the total). We found that 7 of these had expired when we fetched the Whois data and we also determined that 43 were no longer hosting adult material.[16]

We analysed the Whois data for the remaining 3434 domains. A random sample of 40 domains was selected from the 1770 domains with apparently valid phone numbers for the registrant and attempts were made to make contact.

The overall results for WP6.5, in standard format, were:

no Whois	98		
no phone		**92**	2.8%
phone		**1770**	53.1%
privacy	118		
proxy	1356		
privacy+proxy		**1474**	44.2%
	TOTAL	**3336**	

phone call results:	measured	inferred
invalid phone number/does not connect	10	38
number is not answered	5	114
inconclusive call / answering machine	19	113
number does not work to reach registrant	0	9
number works to reach registrant	6	33
TOTAL	40	307

[16] We visited a few of the websites in this category to confirm our view that these sites do not often change their general nature, but we did not check the content of all of them. The 43 domains that we identified as no longer hosting adult material resulted from the domains being described in the *whois* data as being for sale, or because they were hosted on well-known 'domain parking' sites.

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	1474	44.2%
no phone number in Whois	92	2.8%
invalid phone number/does not connect	219	6.6%
number is not answered	657	19.7%
inconclusive call / answering machine	651	19.5%
number does not work to reach registrant	52	1.6%
number works to reach registrant	190	5.7%

It will be noted that there were a non-trivial number (98) domains for which we were unable to obtain Whois data of which 97 were in .com and .net. The .com and .net gTLDs employ a two-level Whois system. In all cases a valid response was received at the first level from whois.crsnic.net, but in 66 cases there was no response at all from the Whois server that whois.crsnic.net indicated should be consulted next. In the other 31 cases there was a response from the second level Whois server, but this response was to the effect that no data was available – implying that the first level response was incorrect.

12.6 WP6.6 Typosquatted domains

We consider the 1227 domain names from work package WP4 (typosquatting) for which we identified any candidate typosquatting domains. These domains are, by very definition, being used by extremely popular websites and they were not registered in the furtherance of any criminal activity.

Since WP4 only considered .com domains, all of these domains are in .com.

A number of these sites were included in other work packages (where they tend to be described as "third party" sites) hence although we only randomly selected 40 apparently valid numbers to call, we did in the course of the whole project, call 50 of them.

The results of WP6.6, in standard format were:

no *whois*	0		
no phone		**81**	6.6%
phone		**911**	74.2%
privacy	48		
proxy	187		
privacy+proxy		**235**	19.2%
	TOTAL	**1227**	

phone call results:	measured	inferred
invalid phone number/does not connect	11	21
number is not answered	5	11
inconclusive call / answering machine	10	17
number does not work to reach registrant	2	4
number works to reach registrant	22	34
TOTAL	50	87

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	235	19.2%
no phone number in Whois	81	6.6%
invalid phone number/does not connect	220	17.9%
number is not answered	115	9.4%
inconclusive call / answering machine	178	14.5%
number does not work to reach registrant	42	3.4%
number works to reach registrant	356	29.0%

13. WP7 'Domains appearing in email spam (SURBL domains)'

The SURBL organisation (the name originates from the term "Spam URI Realtime Blocklist") maintains a database that can be used for blocking messages on the basis of the URLs found within the message.

SURBL provided us with a feed of their "multi-surbl-list" which (at the time we sampled it) combined six specialist lists:

SC: message-body web sites processed from SpamCop URI reports, also known as "spamvertised" web sites.

WS: records created from Bill Stearns' SpamAssassin ruleset sa-blacklist plus many other data sources.

OB: records created from data provided by Outblaze, who analyse message bodies and process user reports.

AB: records fromAbuseButler for Spamvertised Sites which have been most often reported over the previous seven days.

PH: data from multiple sources that identify phishing URLs or sites that host malware.

JP: data generated by running Joe Wein's jwSpamSpy over unsolicited messages; yielding domains providing fulfilment for the advertised product, and/or malware.

As can be seen, the SURBL list contains a mixture of domains registered for criminal purposes, domains for websites that have been compromised, and domains owned by third parties who provide services such as URL shorteners and web hosting.

13.1 Raw data for this work package

We recorded the 28264 domains (and 1184 IP addresses) listed by SURBL over a 7 day period (18 July 2012 to 24 July 2012). The domains were in 98 different TLDs of which the top 30 were:

com	**11946**	**42.3%**	pl	230	0.8%	cc	27	0.1%
info	**5528**	**19.6%**	**biz**	**229**	**0.8%**	ar	26	0.1%
net	**2189**	**7.7%**	at	145	0.5%	be	26	0.1%
tk	2084	7.4%	br	145	0.5%	cn	24	0.1%
ru	1902	6.7%	nl	83	0.3%	ws	22	0.1%
org	**1064**	**3.8%**	de	68	0.2%	mobi	21	0.1%
in	876	3.1%	ua	65	0.2%	cl	20	0.1%
us	543	1.9%	es	63	0.2%	co	20	0.1%
eu	279	1.0%	jp	47	0.2%	fr	20	0.1%
uk	250	0.9%	it	34	0.1%	au	19	0.1%

Between them, the five gTLDs being studied cover 20956 domains (74.1% of the total).

13.2 Results

There were 35 domains which were operated by third parties – where, for example, SURBL listed malicious subdomains. There are so few of these that no useful conclusions can be drawn from this data and we do not consider them any further, but just analyse the remaining 20921 domains.

The left hand table presents the results from processing the Whois data, whether a privacy or proxy service has been used, and if not, whether there is an apparently valid phone number for the registrant.

The right hand table gives the results from making a phone call to a sample of the (non-privacy non-proxy service) registrants when an apparently valid phone number was present in the Whois. The first column of this table shows the 'measured' results, and the second column shows the results we have 'inferred' by assuming that we would get the same result for all the other domains in all of our samples (from all work packages) which had the same phone number as one that we called.

The lower right hand table shows the effect of scaling up the inferred results to cover the whole of each category of domains (we need to do this scaling to include the "no phone number" category so that we can ascertain the percentage of calls that are likely to result in particular outcomes).

no *whois*	157		
no phone		**220**	1.1%
phone		**11379**	54.8%
privacy	74		
proxy	9091		
privacy+proxy		**9165**	44.1%
	TOTAL	**20764**	

phone call results:	measured	inferred
invalid phone number/does not connect	98	1311
number is not answered	11	761
inconclusive call / answering machine	60	4900
number does not work to reach registrant	20	542
number works to reach registrant	12	142
TOTAL	201	7656

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	9165	44.1%
no phone number in *whois*	220	1.1%
invalid phone number/does not connect	1949	9.4%
number is not answered	1131	5.4%
inconclusive call / answering machine	7283	35.1%
number does not work to reach registrant	806	3.9%
number works to reach registrant	211	1.0%

It will be noted that there were a non-trivial number (157) domains for which we were unable to obtain Whois data – in 14 cases this was because the domain no longer existed in the TLD database. All of the rest of the domains were in .com or .net, which both employ a two-level Whois system. In every case a valid response was received at the first level from whois.crsnic.net, but there was no response at all from the Whois server that whois.crsnic.net indicated should be consulted next.

Discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

14. WP8 'Domains associated with malware (StopBadware domains)'

StopBadware is a non-profit anti-malware organization which works to prevent, mitigate, and remediate "badware" websites – websites that serve up viruses, spyware, scareware, and other badware.[17]

14.1 Raw data for this work package

We were provided with a list of 41878 URLs by the StopBadware project – which was one week's worth of new badware URLs.

These URLS used 20331 distinct domains and 117 IP addresses. There were no less than 146 different TLDs represented, of which the most prevalent 33 were:

com	**8427**	**41.4%**
ru	1792	8.8%
net	**1075**	**5.3%**
org	**887**	**4.4%**
de	689	3.4%
pl	679	3.3%
tk	464	2.3%
info	**413**	**2.0%**
it	413	2.0%
br	341	1.7%
nl	336	1.7%

ua	336	1.7%
uk	313	1.5%
eu	234	1.2%
au	171	0.8%
fr	169	0.8%
es	146	0.7%
in	144	0.7%
ro	142	0.7%
ar	130	0.6%
biz	**125**	**0.6%**
cz	121	0.6%

pro	113	0.6%
tr	112	0.6%
cn	104	0.5%
ca	100	0.5%
be	91	0.4%
ch	87	0.4%
hu	86	0.4%
dk	81	0.4%
se	81	0.4%
gr	77	0.4%
us	76	0.4%

Between them, the five gTLDs being studied cover 10927 domains (53.4% of the total).

14.2 Results

There were 94 domains which were operated by third parties.[18] There are so few of these, relatively speaking, that no useful conclusions can be drawn from this group and we do not consider them any further, but just analyse the remaining 10833 domains.

The left hand table presents the results from processing the Whois data, whether a privacy or proxy service has been used, and if not, whether there is an apparently valid phone number for the registrant.

The right hand table gives the results from making a phone call to a sample of the (non-privacy non-proxy service) registrants when an apparently valid phone number was present in the Whois. The first column of this table shows the 'measured' results, and the second column shows the results we have 'inferred' by assuming that we would get the same result for all the other domains in all of our samples (from all work packages) which had the same phone number as one that we called.

The lower right hand table shows the effect of scaling up the inferred results to cover the whole of each category of domains (we need to do this scaling to include the "no phone number" category so that we can ascertain the percentage of calls that are likely to result in particular outcomes).

[17] https://www.stopbadware.org/
[18] We use the same definition of "third party" as we did in work package WP1t (see Section 7.3).

no Whois	114		
no phone		**553**	5.2%
phone		**7983**	74.5%
privacy	150		
proxy	2033		
privacy+proxy		**2183**	20.4%
	TOTAL	**10719**	

phone call results:	measured	inferred
invalid phone number/does not connect	53	186
number is not answered	23	26
inconclusive call / answering machine	50	104
number does not work to reach registrant	9	19
number works to reach registrant	66	254
TOTAL	201	589

Scaling up the inferred values to the whole dataset:

uses privacy or proxy service	2183	20.4%
no phone number in Whois	553	5.2%
invalid phone number/does not connect	2521	23.5%
number is not answered	352	3.3%
inconclusive call / answering machine	1410	13.2%
number does not work to reach registrant	258	2.4%
number works to reach registrant	3443	32.1%

It will be noted that there were a non-trivial number (114) domains for which we were unable to obtain Whois data – in 94 cases this was because the domain no longer existed in the TLD database or was in the process of expiring. The remaining 20 domains were in .com, which employs a two-level Whois system. In every case a valid response was received at the first level from whois.crsnic.net, but there was no response at all from the Whois server that whois.crsnic.net indicated should be consulted next.

Discussion of these results, and the extent to which they are statistically significant, can be found later in the report, once all of the work packages have been fully described.

15. WP9 'Domains subject to the UDRP process'

We considered a sample of domain names that have been subject to ICANN's Uniform Domain Name Dispute Resolution Policy (UDRP). As ICANN explains:[19]

> Under the policy, most types of trademark-based domain-name disputes must be resolved by agreement, court action, or arbitration before a registrar will cancel, suspend, or transfer a domain name. Disputes alleged to arise from abusive registrations of domain names (for example, cybersquatting) may be addressed by expedited administrative proceedings that the holder of trademark rights initiates by filing a complaint with an approved dispute-resolution service provider.

At present there are four approved dispute resolution providers, and we constructed our samples as follows:

WIPO

We considered the 185 cases decided in January 2013.

National Arbitration Forum

We considered the 154 cases decided in January 2013.

The Czech Arbitration Court Arbitration Center for Internet Disputes

We considered the 3 cases decided in January 2013.

Asian Domain Name Dispute Resolution Centre

Unlike the other dispute resolution providers, this organization does not list cases by decision date, so we considered the 12 cases commenced in November 2012 and not subsequently withdrawn, all of which had been decided by the time that we performed our analysis.

A number of these cases involved multiple domain names, which were believed to have a common registrant. In these cases we analysed the first domain of the set which comes from the five gTLDs that we are considering in this report. If there was no such domain then we just considered the first domain that was listed in the decision document. With this definition, the 354 cases concerned domains in 19 TLDs (noting that many TLDs do not employ the UDRP mechanisms we have studied):

com	**274**	**77.4%**
net	**23**	**6.5%**
org	**12**	**3.4%**
biz	**8**	**2.3%**
info	**6**	**1.7%**
co	5	1.4%
es	4	1.1%

mx	4	1.1%
mobi	3	0.8%
nl	3	0.8%
ro	3	0.8%
au	2	0.6%
asia	1	0.3%

ma	1	0.3%
me	1	0.3%
pro	1	0.3%
tel	1	0.3%
tv	1	0.3%
xxx	1	0.3%

[19] http://www.icann.org/en/help/dndr/udrp

We examined the Whois data for the five gTLDs we are studying (323 domains, 91.2%) using a commercial provider of archived Whois responses to determine what data was being served immediately prior to the commencement of UDRP proceedings.

No attempt was made to make contact with any of the domain registrants in this category. This was partly we would have been processing data many months after it was current but mainly because in the vast majority of cases the domain was transferred from the original registrant and that person might not be minded to be helpful in responding to our survey.

Presenting the results in standard form gives this table:

no Whois	3		
no phone		**15**	4.7%
phone		**178**	55.6%
privacy	4		
proxy	123		
privacy+proxy		**127**	39.7%
	TOTAL	**320**	

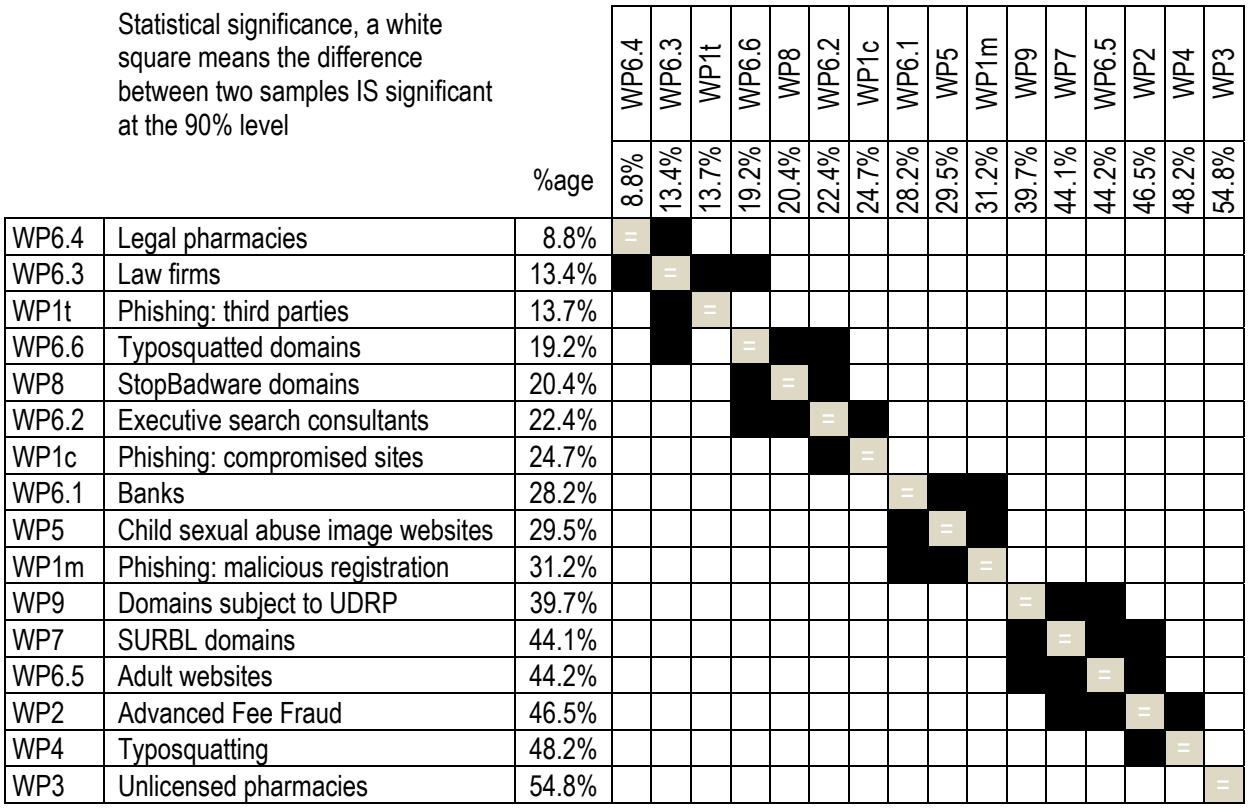
Further discussion of these results, and the extent to which they are statistically significant, can be found in the next section, now that all of the work packages have been fully described.

16. Usage of privacy and proxy services

We now consider the usage of privacy and proxy services in each of the categories from each of the work packages, collating all of the data from earlier in the report.

	Work package	Total	Privacy		Proxy			%age	Maliciously registered?
WP6.4	Legal pharmacies	251	16	+	6	=	22	8.8%	no
WP6.3	Law firms	112	14	+	1	=	15	13.4%	no
WP1t	Phishing: third parties	263	2	+	34	=	36	13.7%	no
WP6.6	Typosquatted domains	1227	48	+	187	=	235	19.2%	no
WP8	StopBadware domains	10719	150	+	2033	=	2183	20.4%	some
WP6.2	Executive search consultants	183	24	+	17	=	41	22.4%	no
WP1c	Phishing: compromised sites	2121	37	+	487	=	524	24.7%	no
WP6.1	Banks	1405	352	+	44	=	396	28.2%	no
WP5	Child sexual abuse image websites	597	19	+	157	=	176	29.5%	yes
WP1m	Phishing: malicious registration	449	29	+	111	=	140	31.2%	yes
WP9	Domains subject to UDRP	320	4	+	123	=	127	39.7%	some
WP7	SURBL domains	20764	74	+	9091	=	9165	44.1%	mostly
WP6.5	Adult websites	3336	118	+	1356	=	1474	44.2%	no
WP2	Advanced Fee Fraud	651	21	+	282	=	303	46.5%	yes
WP4	Typosquatting	27007	378	+	12642	=	13020	48.2%	yes
WP3	Unlicensed pharmacies	765	11	+	408	=	419	54.8%	yes

The differences between samples are only statistically significant as shown:[20]



[20] We use a χ^2 test to compare each pair of samples.

What this graph illustrates is that, generally, when the usage of privacy and proxy services in pairs of categories differ by more than three percentage points then this difference is statistically significant at the 90% level.

However, this generalisation is NOT the case for comparisons involving WP6.2, WP6.3 and WP9 because these categories all have relatively small sample sizes. The percentage of privacy and proxy usage that we saw for these is NOT very different (at a 90% significance level) from that of several other work packages.

That is (to take one example), there is NOT a significant difference between WP6.3's result (13.4%) and WP6.4 (8.8%), WP1t (13.7%) and WP6.6 (19.2%). However, because of the larger sample sizes involved, there IS a significant difference between WP1t (13.7%) and WP6.6 (19.2%).

It will be noted that usage of privacy services is usually only a small fraction of the usage of proxy services, with the exception of WP6.1, WP6.2, WP6.3 and WP6.4. When we analysed the domains in these work packages, we found that many were registered with Network Solutions using this registrar's "Private Domain Registration" privacy service. We observed that most other registrars proposed the use of a proxy service to registrants who wished to conceal details about themselves.

We now summarise the evidence we have of linkage between malicious registrations of domains and the usage of privacy or proxy services.[21]

	Work package	Maliciously registered?	Usage of privacy or proxy services
WP6.4	Legal pharmacies	no	low
WP6.3	Law firms	no	low
WP1t	Phishing: third parties	no	low
WP6.6	Typosquatted domains	no	average
WP8	StopBadware domains	some	average
WP6.2	Executive search consultants	no	average
WP1c	Phishing: compromised sites	no	average
WP6.1	Banks	no	high
WP5	Child sexual abuse image websites	yes	high
WP1m	Phishing: malicious registration	yes	very high
WP9	Domains subject to UDRP	some	very high
WP7	SURBL domains	mostly	very high
WP6.5	Adult websites	no	very high
WP2	Advanced Fee Fraud	yes	extremely high
WP4	Typosquatting	yes	extremely high
WP3	Unlicensed pharmacies	yes	extremely high

The table clearly shows a correlation in that maliciously registered domains generally have a higher usage of privacy and proxy services – but this correlation is not universal in that banks are above average users of these services, as are adult websites.

[21] We call 15% to 25% 'average' as this range covers the rate of privacy or proxy service use of the various NORC studies, as well as the two categories WP6.6 and WP1c each of which has a wide range of different types of site that have been registered for legitimate purposes. Below this range we deem 'low' and above this range we deem 'high'. We use the adjective 'very' if the value is more than 10% above or below our average band and the adjective 'extremely' for more than 20%.

16.1 Comparison with the NORC privacy and proxy service data

We can also compare our results with that of the recent, May 2013, NORC study:[22]

privacy or proxy registration	320	20.0%
normal registration	1280	80.0%
TOTAL	1600	

The NORC study sampled domains from the complete set of those that are registered so it is appropriate to consider error bounds – the size of the population they sampled means that their results are, at the 90% level of significance we are using in this report, ± 1.6%.

We find, using a χ^2 test, that there IS a significant difference (higher or lower) with all our categories except WP6.6 (typosquatted domains, 19.2%), WP8 (StopBadware domains, 20.4%) and WP6.2 (Executive search consultants, 22.4%).

For their analysis of privacy and proxy services NORC took samples from all registered domains whereas in every category we consider there is a website involved – and not all domains are associated with websites.

NORC found that 416 (26.0%) domains from their sample had no online website presence and 328 (20.5%) domains were 'parked'. That is, only just over half of their sample is directly comparable with the types of domains that we were considering. That said, Table A-1 shows that if we exclude these two types of domain then they measured the usage of privacy or proxy services at 17.3% (a lower figure because they measure the privacy or proxy service usage for parked domains to be 30%).

However, NORC also found that domains registered to legal persons had a 15.1% usage of privacy or proxy services – and inspection of our figures from WP6, which is considering activities generally undertaken by legal persons mainly shows lower usages of privacy and proxy services. That is, our findings are not grossly inconsistent with NORCs results.

16.2 Overall conclusion on privacy and proxy service usage

We initially set out to consider the two hypotheses

"A significant percentage of the domain names used to conduct illegal or harmful Internet activities are registered via privacy or proxy services to obscure the perpetrator's identity".

and

"The percentage of domain names used to conduct illegal or harmful Internet activities that are registered via privacy or proxy services is significantly greater than the percentage of domain names used for lawful Internet activities that employ privacy or proxy services."

We conclude that the first hypothesis is supported by our data, with 29% or more of domains associated with illegal or harmful activities being registered via privacy or proxy services.

However, the second hypothesis is only partly correct. Privacy and proxy usage for the categories of unlawful activities sampled in our study ranges from 20% (WP8, StopBadware domains) to 55% (WP3, unlicensed pharmacies). This range overlaps considerably with the equivalent percentages for the sampled lawful and harmless activities: 9% (WP6.4, legal pharmacies) to 44% (WP6.5, adult websites).

[22] http://gnso.icann.org/en/issues/whois/registrant-identification-summary-23may13-en.pdf

17. Validity of contact phone numbers

For most work packages (not WP5 or WP9) we made phone calls to samples of domain registrants for whom we had apparently valid contact numbers and this table first summarises the results we already presented. Since these values are scaled up from our samples, we now calculate the error bounds. We are assuming that calling the same number about a further domain would yield the same result as a call that was actually made.[23]

Sample		Total	no phone number in *whois*	invalid phone number / does not connect	number is not answered	inconclusive call / answering machine	number does not work to reach registrant	number works to reach registrant
WP1c	Phishing: compromised sites	2121	5.1%	31.0%	11.8%	2.7%	0.9%	23.7%
WP1m	Phishing: malicious registration	449	5.3%	48.4%	5.3%	0.4%	7.5%	1.8%
WP1t	Phishing: third parties	263	4.2%	31.2%	17.0%	1.1%	0.5%	32.3%
WP2	Advanced Fee Fraud	651	1.5%	31.4%	3.8%	5.1%	9.4%	2.1%
WP3	Unlicensed pharmacies	765	0.1%	33.4%	5.3%	2.2%	3.5%	0.7%
WP4	Typosquatting	27007	4.0%	4.8%	5.0%	16.8%	10.6%	10.6%
WP6.1	Banks	1405	7.8%	5.4%	6.5%	33.7%	3.3%	15.2%
WP6.2	Executive search consultants	183	5.5%	5.3%	14.1%	15.8%	3.5%	33.4%
WP6.3	Law firms	112	10.7%	9.5%	7.6%	34.2%	0.0%	24.7%
WP6.4	Legal pharmacies	251	6.0%	7.6%	0.0%	49.5%	4.6%	23.6%
WP6.5	Adult websites	3336	2.8%	6.6%	19.7%	19.5%	1.6%	5.7%
WP6.6	Typosquatted domains	1227	6.6%	17.9%	9.4%	14.5%	3.4%	29.0%
WP7	SURBL domains	20764	1.1%	9.4%	5.4%	35.1%	3.9%	1.0%
WP8	StopBadware domains	10719	5.2%	23.5%	3.3%	13.2%	2.4%	32.1%
Error bounds (at 90% confidence level):								
WP1c	Phishing: compromised sites	2121	exact	±14.8%	±8.1%	±3.7%	±2.1%	±11.3%
WP1m	Phishing: malicious registration	449	exact	±5.5%	±3.2%	±0.8%	±4.4%	±1.6%
WP1t	Phishing: third parties	263	exact	±13.6%	±8.6%	±0.9%	±0.6%	±12.4%
WP2	Advanced Fee Fraud	651	exact	±3.5%	±2.2%	±2.2%	±2.7%	±1.7%
WP3	Unlicensed pharmacies	765	exact	±1.4%	±1.2%	±0.5%	±0.9%	±0.2%
WP4	Typosquatting	27007	exact	±5.5%	±8.5%	±18.3%	±24.7%	±15.0%
WP6.1	Banks	1405	exact	±8.7%	±15.0%	±16.2%	±4.8%	±11.7%
WP6.2	Executive search consultants	183	exact	±5.6%	±9.4%	±9.0%	±4.7%	±11.0%
WP6.3	Law firms	112	exact	±6.3%	±5.7%	±9.5%	±0.0%	±9.0%
WP6.4	Legal pharmacies	251	exact	±6.3%	±0.0%	±18.3%	±5.4%	±13.8%
WP6.5	Adult websites	3336	exact	±9.4%	±35.0%	±23.2%	±4.6%	±10.8%
WP6.6	Typosquatted domains	1227	exact	±11.9%	±12.0%	±9.9%	±5.6%	±13.9%
WP7	SURBL domains	20764	exact	±5.3%	±4.4%	±8.2%	±3.5%	±1.2%
WP8	StopBadware domains	10719	exact	±15.6%	±2.1%	±8.1%	±2.2%	±19.1%

[23] This complicates the estimation of error bounds, and the method we have used is that outlined in Snedecor & Cochran, *Statistical Methods*, Edition 6, page 515. That is, for each inferred proportion p measured from our sample of phone calls we calculate the standard error using the formula:

$$s = \sqrt{\frac{1}{n(n-1)}\sum\left\{\left(\frac{m_i}{\bar{m}}\right)^2 (p_i - p)^2\right\}(1-\Phi)}$$

where n is the number of phone numbers sampled, and p_i is either 1 or 0 depending on whether the response to that phone is in the portion of interest, m_i is the number of domains registered with the i^{th} phone number, $\bar{m}$ is the average number of registered domains per phone number, and Φ is the sampling fraction (the proportion of work package domains for which a phone call was made).

Having calculated *s*, the standard deviation of the proportions, we assume a normal distribution and apply a two-tailed test.

The error ranges which we document show the impact of relatively small sample sizes (we only made 40 calls in some of the categories) along with the uncertainty caused for some work packages (notably WP4) where there were many hundreds of domains with the same contact phone number… so the result of one or two phone calls could make a substantial difference to the measured outcome.

17.1 Does a phone call reach the domain registrant ?

We now consider whether or not we can reach the domain registrant (this is extracted from the previous table and presented in this new table for clarity). This gives extremely striking results:

		Sample size	Registrant contacted	Error range	Maliciously registered?
WP6.2	Executive search consultants	183	33.4%	±11.0%	no
WP1t	Phishing: third parties	263	32.3%	±12.4%	no
WP8	StopBadware domains	10719	32.1%	±19.1%	some
WP6.6	Typosquatted domains	1227	29.0%	±13.9%	no
WP6.3	Law firms	112	24.7%	±9.0%	no
WP1c	Phishing: compromised sites	2121	23.7%	±11.3%	no
WP6.4	Legal pharmacies	251	23.6%	±13.8%	no
WP6.1	Banks	1405	15.2%	±11.7%	no
WP4	Typosquatting	27007	10.6%	±15.0%	yes
WP6.5	Adult websites	3336	5.7%	±10.8%	no
WP2	Advanced Fee Fraud	651	2.1%	±1.7%	yes
WP1m	Phishing: malicious registration	449	1.8%	±1.6%	yes
WP7	SURBL domains	20764	1.0%	±1.2%	mostly
WP3	Unlicensed pharmacies	765	0.7%	±0.2%	yes

We find that domain registrants who are using their domains for most of the lawful purposes that we studied can be reached between 15.2% and 33.4% of the time (one time in six, to one time in three).

However, where maliciously registered domains are being used for criminal purposes, the phone number provided in the Whois reaches the domain registrant 2.1% of the time at the very most. That is, no more than one time in fifty.[24]

The two exceptions to this general result are typosquatting – which, as we have already discussed, is not a criminal matter – and operating adult websites, which although some people may disapprove of the activity, will be lawful in the jurisdictions where the websites are hosted.

Reaching the domain registrant by phone is only possible 5.7% (± 10.8%) of the time for the adult website domains. In contrast, registrants of the typosquatting domains – those domains that have been registered solely because of their similarity to the major brands – can be reached by phone almost twice as often, 10.6% (±15.0%) of the time.

[24] Strictly, we should apply the error ranges, and then say that we're 90% sure that we can reach most of the lawful registrants between 3.5% and 44.4% of the time (the lower bound being 9.8% if we exclude banks), and the malicious domain registrants can be reached at most 3.8% of the time.

17.2 Is it impossible to make a phone call to reach the registrant ?

The large error bounds for our results in many of the categories (caused by many domains having the same contact details, and limited numbers of survey calls being made) means that these conclusions, although highly suggestive of some underlying truth, might possibly have arisen by chance.

However, many of the phone calls that we made that didn't succeed didn't entirely fail either. For example, calls to phone numbers associated with banks, law firms and legal pharmacies often went direct to voicemail. This leads us to looking at the data the other way around – not "can we reach the registrant by phone" but rather "is it entirely impossible, using just Whois data to make a phone call to the party using the domain ".

The gap between these two ways of looking at the data occurs when we may or may not be able to reach the domain user by phone. Just because we failed to reach the relevant person to ask our survey question does not mean that someone else, with a rather different message to deliver, might not succeed. In particular, our chosen methodology meant that when we encountered an answering machine or voicemail system we treated this as an inconclusive call – we felt it would be unreasonable to expect people to call back to answer our survey; but issues of more immediate importance to them could well turn out differently.

For the analysis of "impossible to reach by phone" we sum the following cases:

- Having a registration that uses a privacy or proxy service

  When either a privacy or proxy registration is chosen, the domain registrant or licensee's phone number is not made public.

- No valid phone number in the Whois

  This is entirely straightforward – no 'apparently valid' phone number was present for the domain registrant. This category includes numbers which are too short to be valid, consist of all 9s, all 0s or which have invalid area codes.[25]

- Phone number in the Whois fails to work (category #1 in section 6.5)

  This is essentially the same situation as the previous category, but our simple rules for invalidity were not triggered – we only discovered that the number was invalid when an attempt was made to call it. This category does *not* include calls where the number rings and rings, is a cellphone that the network states is not reachable, or any of the cases where the phone is answered, whether by a human or a voicemail system.

- Person who is reached denies registering the domain (category #4 in section 6.5)

  The number is valid, but the call reaches someone else who denies that the phone number is suitable for reaching the person who is recorded as registering the domain – or the number reaches the correct person, but they deny having registered the domain.

[25] We would caution that our methodology only considers whether we have a phone number specifically for the registrant. Some registrars who provide a registrant's name and address never provide an accompanying email address or phone number, but they do provide a full set of details for administrative, technical and billing contacts. It is often the case that one of the other sets of details has the same name and address as provided for the registrant and we then deem the phone number as being that of the registrant. Unfortunately, this creates a bias against larger companies where it is common to see details of hosting companies or website designers as technical, billing or administrative contacts.

The results of this 'impossible to call' analysis are given in the following table:[26]

		sample size	uses privacy / proxy service	no phone number in *whois*	invalid phone number / does not connect	number does not work to reach registrant	not possible to phone the registrant	error bounds	maliciously registered?
WP6.4	Legal pharmacies	251	8.8% +	6.0% +	9.5% +	0.0% =	24.2%	1.2%	no
WP6.3	Law firms	112	13.4% +	10.7% +	9.5% +	0.0% =	33.6%	0.7%	no
WP6.2	Executive search consultants	183	22.4% +	5.5% +	5.3% +	3.5% =	36.7%	0.9%	no
WP6.1	Banks	1405	28.2% +	7.8% +	5.4% +	3.3% =	44.6%	0.9%	no
WP6.6	Typosquatted domains	1227	19.2% +	6.6% +	17.9% +	3.4% =	47.1%	2.8%	no
WP1t	Phishing: third parties	263	13.7% +	4.2% +	31.2% +	0.5% =	49.6%	4.3%	no
WP8	StopBadware domains	10719	20.4% +	5.2% +	23.5% +	2.4% =	51.4%	4.4%	some
WP6.5	Adult websites	3336	44.2% +	2.8% +	6.6% +	1.6% =	55.1%	0.8%	no
WP7	SURBL domains	20764	44.1% +	1.1% +	9.4% +	3.9% =	58.5%	0.8%	mostly
WP1c	Phishing: compromised sites	2121	24.7% +	5.1% +	31.0% +	0.9% =	61.7%	5.3%	no
WP4	Typosquatting	27007	48.2% +	4.0% +	4.8% +	10.6% =	67.7%	3.9%	yes
WP2	Advanced Fee Fraud	651	46.5% +	1.5% +	31.4% +	9.4% =	88.9%	2.4%	yes
WP3	Unlicensed pharmacies	765	54.8% +	0.1% +	33.4% +	3.5% =	91.8%	1.2%	yes
WP1m	Phishing: malicious registration	449	31.2% +	5.3% +	48.4% +	7.5% =	92.5%	7.5%	yes

As can be seen, it is impossible to consider reaching the registrant of the domains for the lawful businesses we studied in proportions that vary from 24.2% (WP6.4, legal pharmacies) to 55.1% (WP6.5, adult websites) – with the compromised phishing website domains (WP1c) very slightly higher than this at 61.7%, possibly because this last category contains many websites that are operated by individuals or micro-businesses.

The registrants in the WP7 (SURBL) and WP8 (StopBadware) categories (which both contain mixtures of domain names – some of which were maliciously registered explicitly for criminal purposes), are unreachable in the proportions 58.5% and 51.4%.The proportion of typosquatting domain registrants (WP4) who cannot be reached by phone is 67.7%.

The domains from the entirely criminal categories WP2 (advanced fee fraud), WP3 (unlicensed pharmacies) and WP1m (maliciously registered phishing domains) are registered by people who are unreachable by phone in 88.9% to 92.5% of cases. These figures are remarkably similar – even though there are considerable variations to be seen in whether or not privacy or proxy services are used.

It should also be noted that the anecdotal evidence from the Internet Watch Foundation for WP5 (child sexual abuse image websites) is that 100% of the people who create these sites cannot be contacted, although only 29.5% of the WP5 domain registrants used privacy or proxy services.

[26] The error bounds here relate to the sampling and scaling actions which give the numbers in the second two columns of the sum (see footnote 23 for the statistical test that is employed). The first two numbers in the sum are 'exact' so they do not contribute to the error bounds. The significance of the differences between the values of relevant categories is discussed in the text.

18. What is not in this study

ICANN's original specification for this study[27] was changed during the negotiation process between NPL and ICANN and this changed specification formed the contractual basis of the study actually undertaken. However, as is apparent from the feedback described in Appendix B, some of the respondents only appear to be aware of the original study specification, not the final version agreed between NPL and ICANN.[28]

This section briefly sets out our reasoning for proposing to ICANN that we should not address the other topics set out in ICANN's original specification for this study, or that we should address them in other ways than was originally envisaged. It may also be helpful to consult Appendix B where we elaborate on a few of these points.

We have tried extremely hard to obtain unbiased lists of domains for all of our work packages, by having some objective way of obtaining the data. Where we did take feeds of "bad things" we concentrated on ensuring that they were as complete as possible, so we used very well-known sources with large numbers of contributors (SURBL, APWG etc.) or we looked for authoritative sets of data that covered an entire topic (such as from the IWF).

For many of the categories we did not wish to study we were, and remain, unaware of any source of data which could be said to be relatively unbiased in coverage and global in scope. Without having a sufficiently large set of unbiased data it is simply not possible to study a topic in a scientific manner

For some of the topics that we did not wish to cover we believe that the activity is very similar to one of the topics we did study and we explain the reasons for our belief below.

If we had studied any of these topics then, in our view, none of the results that we would have been able to obtain would have been particularly useful in strengthening our tests of the hypotheses about domain name registration.

'Spam' It was suggested that that "live-feeds" from several major real-time Domain Name System Blacklists (DNSBLs) could be used to generate a subsample of spam sender IP addresses/ranges and associated unique domain names. We consider that this is an inappropriate way of analysing activity on today's Internet. The DNSBLs generally contain the IP addresses of botnet machines, since that is the main way in which email spam is distributed. The small numbers of DNSBLs that list domains, such as SURBL (whose data we used in WP7), are not concerned with IP address ranges, but with checking the reputation of the website URLs mentioned within the email spam text.

It should perhaps be noted in passing that a completely different Whois service is used for looking up IP address allocation information. This system, operated by the Regional Internet Registries (RIPE ARIN etc.) has nothing to do with Whois system for domain names.

'Malware' The ICANN specification suggested examining specific sources of malware domains, but these are, to a very great extent, lists of domains associated with particular malware instances (such as Koobface, Zeus) etc. or activity detected by particular methodologies such as studying botnet C&C traffic. This would have meant that in effect we

[27] http://gnso.icann.org/issues/whois/whois-proxy-abuse-study-18may10-en.pdf
[28] The GNSO Council's decision was taken on 28 April 2011:
https://community.icann.org/display/gnsocouncilmeetings/Motions+28+April+2011
and is based on an ICANN staff report of 5 October 2010:
http://gnso.icann.org/issues/whois/gnso-whois-pp-abuse-studies-report-05oct10-en.pdf

would have been studying the domain registration choices of particular gangs rather than a particular type of criminality as a whole, and furthermore those gangs are an unknown (because unstudied) proportion of all the activity. Rather than using the sources suggested, we chose to study the malware domains identified by the StopBadware project (in WP8). This list is fairly generic in nature, albeit there is some bias towards sites that can be identified by the 'spidering' activities of search engines.

'Denial-of-service and DNS cache poisoning' We believe that DNS cache poisoning is almost invariably an attack on legitimate domains rather than maliciously registered domains, and we are unaware of any reliable wide-ranging source of data on these fairly rare events. Classic denial-of-service attacks are not dependent on the existence of particular domains and DNS reflection attacks often use legitimate domain data rather than registering domains especially for the purpose.

'Intellectual property theft' ICANN proposed a study should be made of the domains registered in four specific intellectual property issues: "Media Piracy", "Software Piracy", "Trademark Infringement" and "Counterfeit Merchandise".

In WP3 we consider unlicensed pharmacies which (it is often claimed) supply counterfeit merchandise. Studies show that large numbers of domain names are used by numerous affiliates to evade spam filters, and these domains form a substantial fraction of the domains that SURBL records and that we study in WP7. We believe that the people engaged in 'software piracy' operate in broadly similar ways to the pharmacies, and so their activities will also be assessed within WP7.

We consider typosquatting, covered in WP4, to overlap with the trademark infringement topic and to be far more prevalent, improving the relevance to the testing of the study's hypotheses on domain registration.

We did not examine media piracy. The bandwidth requirements of this type of website are substantial, which in turn means that considerable monetary sums will be involved in running such operations. Hence, when attempting to locate the website operators, investigations are far more likely to consider hosting providers, rather than domain name registrants.

'Advance fee fraud' The ICANN document suggested that domains "send solicitation email" but in our experience this email is sent through major webmail systems or from botnets and in neither case is it relevant to study the domain name registrants. In practice, investigations will centre on the 'drop box', the address to which email replies are directed, and these dropboxes are generally sited on webmail systems such as those operated by Yahoo! Google and Microsoft. However, other aspects of Advance Fee Fraud are tackled in WP2.

'Identity theft' The ICANN document suggests that domains "send bait mail associated with online identity theft" (once again, that's not really the case). This topic is covered by WP1 (phishing), WP2 (advance fee fraud, where victim identity information may be sold on) and WP8 (domains in spam). We are not aware of other significant 'identity theft' activity that involves domain names.

'Harassment or stalking' Besides a very significant lack of public data about such activities, investigations once again centre on how email is sent (typically from free webmail systems), or on the identity of people who had set up abusive pages on well-known social networking sites. The IP addresses used would be far more significant than the domain names of purported email contact addresses.

We summarize what we did not wish to study in the following table:

ICANN's suggested topic	Addressed in
Spam	Not fully addressed, see text above; some aspects covered in WP3 and WP7
Phishing (live feed)	WP1
Malware	WP7 & WP8
Denial of service & DNS Cache poisoning	Not fully addressed, see text above; some aspects covered in WP3, WP4 & WP7
Phishing	WP1
Cybersquatting	WP9
Intellectual property theft	Not fully addressed, see text above; some aspects covered in WP3, WP4 & WP7
Media piracy	Not addressed, see text above.
Software piracy	WP7
Trademark infringement	WP4
Counterfeit merchandise	WP3
Money laundering	WP2
Advance fee fraud	WP2
Identity theft	Not addressed, see text above.
'Child pornography'	WP5, where we use the more generally acceptable term of 'child sexual abuse images'
Harassment or Stalking	Not addressed, see text above
Other cybercrime	No additional cybercrime topics were studied

18.1 Analysis that we do not provide

ICANN's original specification for this report set out some very detailed requirements for reporting. In particular, many of the statistics that we report in aggregate were to be broken down by gTLD and by country.

However, in practice, with only a handful of exceptions the datasets that we collected have turned out to be entirely dominated by .com domains, with the other four gTLDs that we studied forming less than 10% of the totals, often very much less.

We do not believe it would be appropriate to provide results for these small subsets since they would be of limited statistical significance and hence any data that we provided would be more likely to mislead rather than to illuminate.

Similarly, the original specification required breakdowns of the use of privacy and proxy services by the declared country in which the domain registrant resides. Once again the sample sizes would make much of this data statistically problematic. Furthermore, it is quite apparent to us that in many of the cases where we have identified that phone numbers are invalid, the rest of the address (and the declared country) is not valid either – so any data we presented would be distorted by this effect.

There is an arguable case for studying whether there are different levels of privacy or proxy service take-up in different countries and in principle the various lawful and harmless categories that we studied would provide this data. However, our methodology has biases towards particular countries (especially the US) for these lawful and harmless activities and there are undoubtedly also biases in the feeds of the different types of unlawful activities from which we sampled. Thus, if we were to present such data, it would be statistically problematic to make any comparisons.

19. Summary and Conclusions

This is one of the largest studies of Whois data ever reported. We processed the registration details in the Whois records of over 70,000 domains. This section summarizes the results that we obtained.

19.1 WP1 (phishing) – the study in a nutshell

The overall results that we obtained can be seen with real clarity in the results of work package WP1 – where we examined domains that had occurred in URLs for phishing pages.

We split this work package into three, since we could analyse the URLs and determine whether the domain was registered by:

- a third party, that is companies which provide services such as hosting or URL shortening – and it so happened that these services were used for criminal purposes;
- a legitimate business (or individual) whose website had been compromised and the phishing web pages added without their knowledge or permission;
- maliciously registered domains, where a registrant intended it to be used for criminal purposes

We found very striking differences between these categories when we considered the usage of privacy or proxy services and also whether we were successful in making contact with the registrant by phone or, conversely, had no hope of doing so:

	using privacy or proxy services		missing or invalid phone number		impossible to contact by phone (all reasons)		contact by phone may, or may not, be possible		contactable by phone		
third party domains	13.7%	+	35.9%	=	49.6%	+	18.1%	+	32.3%	=	100.0%
compromised website domains	24.7%	+	37.0%	=	61.7%	+	14.6%	+	23.7%	=	100.0%
maliciously registered domains	31.2%	+	61.3%	=	92.5%	+	5.8%	+	1.8%	=	100.0%

The people who maliciously registered domains for phishing chose privacy and proxy services somewhat more than people who registered domains for legitimate purposes. However, when a privacy or proxy service was not chosen for a malicious registration a workable contact phone number was seldom given – and even if the number was apparently valid, we almost never managed to make contact with the registrant for our survey.

Conversely, even entirely legitimate 'third party' businesses that provide services to the law-abiding public – and occasionally for malicious purposes – use privacy and proxy services to a certain extent, and for almost half of the domains these businesses use there is no

possibility of using the phone to reach the domain registrant. Of course there are many other ways of making contact with such businesses, and they would doubtless want people to use the information about contact pathways on their websites, rather than consulting Whois.

The compromised website category falls between the two extremes – these domain registrants use privacy and proxy services a quarter of the time (a higher proportion than the NORC study measured). Nearly two thirds of these registrants are impossible to contact by phone, and we reached only a quarter of them for our survey.

19.2 Other categories of criminal or harmful activity

In WP2, we looked at domains registered for advance fee fraud using data collated by the aa419.org project and found a similar result to the maliciously registered phishing domains in WP1 in that 88.9% of domain registrants were not contactable by phone, albeit 46.5% of them chose to privacy or proxy services to achieve this.

In WP3 we examined the domains used for unlicensed pharmacies, finding that 91.8% of the domain registrants were not contactable by phone with 54.8% of them choosing to use privacy or proxy services.

In WP5 we looked at the Whois for domains used for websites containing child sexual abuse images – 29.5% of these use privacy or proxy services and it is widely believed that where contact phone numbers are given for the registrant all of this information is false. That is 100% of these domain registrants cannot be contacted by phone.

19.3 Lawful and harmless activity

We also looked, within WP6, at the domains used for a number of different types of lawful and harmless activity. We found quite large variations in the usage of privacy or proxy services with legal pharmacies (documented on the LegitScript website) at 8.8% and websites listed in the Yahoo! directory as hosting adult material at 44.2% – the latter percentage being somewhat higher than several types of criminal activity.

However, the WP6 domain registrants were, at least to some extent, contactable by phone. Our success rate was highest for executive search consultants (WP6.2) at 33.4%, typosquatted Alexa 3500 companies (WP6.6) at 29.0% and law firms (WP6.3) at 24.7%, but many calls were unanswered, went to voicemail, or we talked to colleagues of the registrant without them being able to assist us in our survey. If all of these call attempts which neither totally failed nor totally succeeded had worked out for us then our success rate would have doubled.

The lowest success rate in WP6 was in making calls to the registrations of domains used for adult websites (WP6.5) where only 5.7% of registrants were reached and 55.1% of the domain registrants were impossible to reach by phone.

19.4 Categories with mixtures of domain registrations

The data from the other work packages is a little harder to interpret. When we look at the results from WP7 (domains listed by SURBL to assist in spam blocking) and WP8 (domains listed by StopBadware which contain various varieties of malware) we find that the WP7 domains have a high usage of privacy and proxy services (44.1%) but WP8 domains use these services less often (20.4%) than the compromised websites from WP1.

Conversely, WP8 domain registrants can be reached by phone 32.1% of the time whereas the figure for WP7 is 1.0%. However, when we look at the "impossible to reach by phone"

measure both WP7 and WP8 have similar figures (58.5% and 51.4%) suggesting that we're seeing similar levels of criminality – both lists are a mixture of maliciously registered domains and legitimately registered domains where a website has been compromised and used to host malicious content.

Significant caution is called for in reading too much into the WP7 data since there are some very high error bounds associated with the WP7 figures. The WP7 data contains a number of groups of domains with the same contact phone number – there are 19 groups of more than 100 domains, and the largest grouping contains 947 domains. These groupings mean that how a handful of registrants respond can substantially affect the results of our survey – and the error bounds reflect this uncertainty.

We suspect that there are some "report inflation" effects occurring in the SURBL data (as we discussed in the detailed account of processing the WP1 data) and in order to best protect the people who use their data they have identified all the domains that could be used to mount an attack rather than just the one that that is currently in use.

Unfortunately, because the datasets we received for WP7 and WP8 only contained hostnames and not full URLs, it was not possible to remove the excess domain names we believe are present in the WP7 data. This is also the reason why we were not able to split these lists to distinguish between maliciously registered domains and legitimate domains. If we had been able to do this, then we would expect to see the sort of differences in the results that we saw in WP1.

19.5 Typosquatting – mixed results

We conclude our review of the work package results by considering WP4 – the typosquatting work package and WP9, the domains involved in UDRP disputes. Almost every dispute in WP9 concerned the type of activity that the WP4 domains are engaged in – with the exception of a handful of cases where brand owners were trying to wrest control of domains away from firms where there was once a close commercial relationship.

As we have noted, typosquatting is a civil matter not a criminal matter, so it might be expected that domain registrants were not quite so cautious about revealing their identity; and conversely that it mattered less anyway – the UDRP process also works with domains that use privacy and proxy services. However, the incentive here for the domain registrant to obscure their identity appears to be the preventing of a brand owner from discerning that a single action could deal with a large number of domains – viz: it's not exactly anonymity that the registrants seek but unlinkability.

The figures here show that privacy and proxy services are used rather more than average (WP4: 48.2%, WP9: 39.7%) but that where domain registrants did provide contact details then in WP7 (we made no phone calls in WP9) we reached the domain registrant for 10.6% of the domains – distinctly more often than the 1%–2% that we measured for domains associated with criminal activities.

However, once again (as in WP7) the data for WP4 has very wide error ranges – many of the domains have the same contact details. Indeed, the original academic paper by Moore and Edelmann found that 63% of typosquatting domains displaying Google ads used just five advert IDs, that is only a handful of people are responsible for a great deal of this activity.

19.6 What can we conclude about the initial hypotheses ?

A final note of caution applies to all of the data we have presented – we have just been looking at domains within biz, com, info, net and org, and for many work packages there are substantial amounts of activity that use other TLDs. We suspect that our results are widely applicable but we have not demonstrated this.

To summarise the whole project and to return at the end to our original hypotheses – we DID find clear evidence that:

"A significant percentage of the domain names used to conduct illegal or harmful Internet activities are registered via privacy or proxy services to obscure the perpetrator's identity".

But, although we did find that it was often true, we DID NOT find that in all cases:

"The percentage of domain names used to conduct illegal or harmful Internet activities that are registered via privacy or proxy services is significantly greater than the percentage of domain names used for lawful Internet activities that employ privacy or proxy services."

Additionally, we learnt that these statements ARE correct:

"When domain names are registered with the intent of conducting illegal or harmful Internet activities then a range of different methods are used to avoid providing viable contact information – with a consistent outcome no matter which method is used.

However, although many more domains registered for entirely lawful Internet activities have viable telephone contact information recorded within the Whois system, a great percentage of them do not."

Domain ownership phone number survey for NPL's ICANN project:

AIM: The aim of the survey is to discover whether the phone number details in WHOIS for the domain registrant (the "domain owner") can be relied upon as a way of reaching that owner. In some cases it is envisaged that the phone numbers will not work, or will reach people whose details match what is in the WHOIS, but who deny having registered the domain.

Note that no further interaction is required beyond establishing whether or not the phone number works to locate the person named in the WHOIS info, and recording whether or not it is the case that they have registered the domain. It is not necessary to speak to the specific person if someone else can confirm that the phone number is a viable way of making contact.

SCHEDULE FOR PHONECALLS:

Numbers which appear to be for businesses:

#1 ring during the morning of a business day

If number fails to connect at all, record this and make no further attempts. Once someone answers then start the protocol set out in the next section. If no answer, then make up to 3 more attempts:

#2 ring during the afternoon of a business day (preferably the same one)
#3 ring during the morning of a different business day
#4 ring during the afternoon of a different business day

Numbers which appear to be for individuals (or where status is unclear):

#1 ring during the late afternoon of a business day

If number fails to connect at all, record this and make no further attempts. Once someone answers then start the protocol set out in the next section. If no answer, then make up to 3 more attempts:

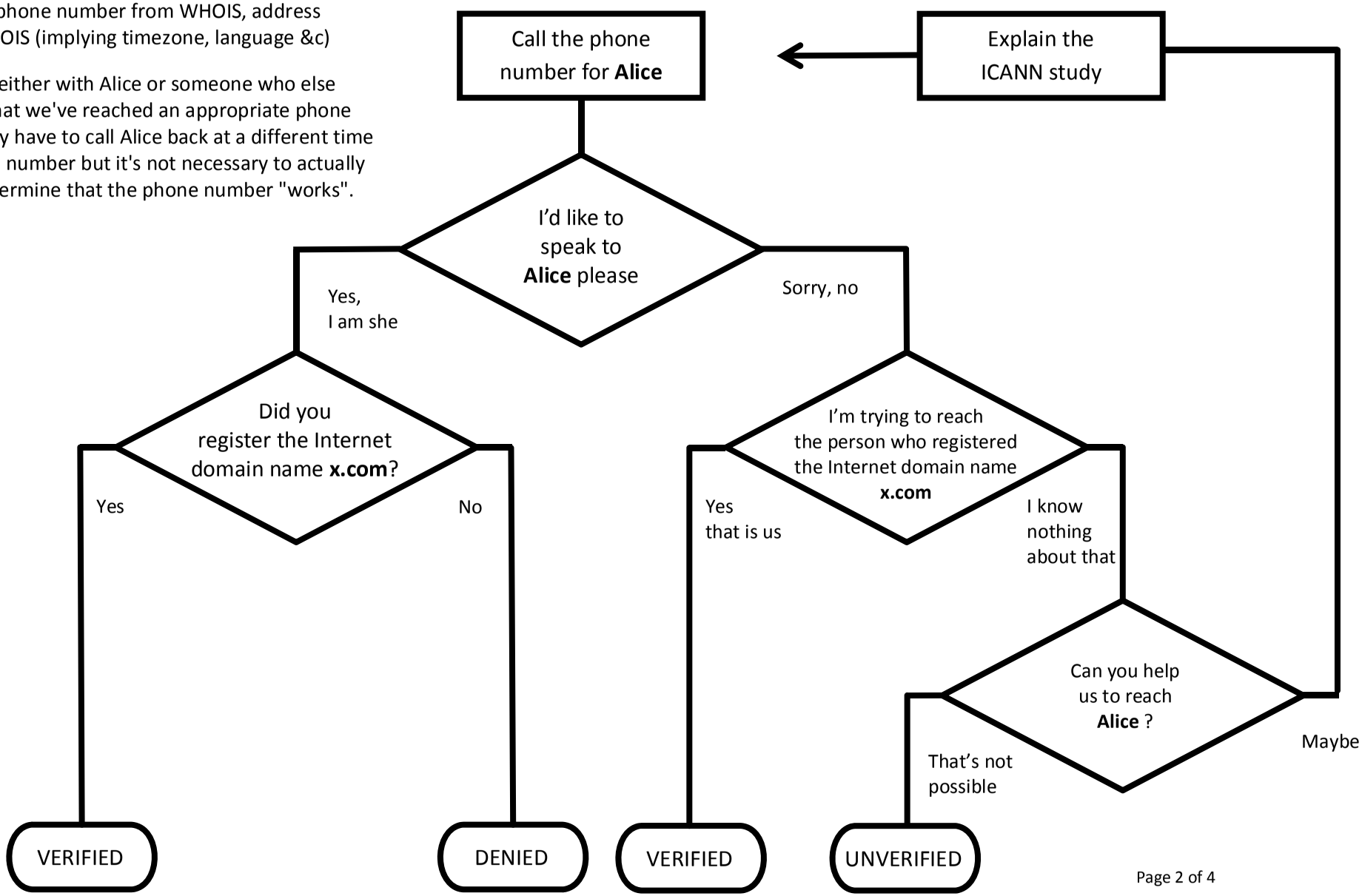
#2 ring during the early evening of another business day
#3 ring during the morning of yet another business day
#4 ring during the mid-evening of a business day

If you reach an answering machine then please adapt the call schedule if the message helps you do this; or if not then continue with the call schedule regardless. If this is a common occurrence we will revise the protocol to address this issue.

PROTOCOL FOR THE PHONE CONVERSATION

You will have: domain name ("x.com"), registrant's name ("Alice"), phone number from WHOIS, address details from WHOIS (implying timezone, language &c)

We wish to talk either with Alice or someone who else who will state that we've reached an appropriate phone number. We may have to call Alice back at a different time or on a different number but it's not necessary to actually reach her to determine that the phone number "works".



USEFUL PHRASES ETC

"Hello, I would like to speak with _____."

"My name is _____ and we're doing a very brief survey to find out if the contact phone numbers recorded for Internet domain names are accurate."

"According to the records, the Internet domain _____ was registered by _____ and the phone number _____ was given."

"The survey is being done by the UK's National Physical Laboratory on behalf of ICANN, the body who looks after the Internet naming systems."

"The domain names were picked entirely at random."

"There is more detail about the survey on the NPL website:"

http://www.npl.co.uk/news/hidden-identities-on-the-web

"If you want to complain that someone has forged your details then you should use the form at:"

http://wdprs.internic.net/

DATA TO BE RECORDED

The phone survey is meant to be very light-weight, and it is intended that recording of results can be done very quickly. Calls should fall into one of the following categories:

Not a legitimate phone number

Number appears legitimate but does not connect

Number unobtainable tone / message

Number rings - no response on any call

Got an answering system message

Spoke to Alice who said that she had registered x.com

Spoke to Alice who said that she had not registered x.com

Spoke to someone other than Alice who confirmed that their husband/wife/company/etc had registed x.com

Spoke to someone other than Alice who denied that Alice existed

Spoke to someone other than Alice who could not help reach Alice

Spoke to someone other than Alice who provided the details for a new call to be made (and then record details of that call as above)

If anyone says anything unusual or noteworthy (most likely when denying that they registered the relevant domain) then we'd appreciate a note of what it was; we'd add such comments, anonymised of course, to the ICANN report to give it some "colour".

Appendix B: Response to critical comments

A draft version of this document was posted by ICANN on 24th September 2013 along with an invitation to the community to comment upon its contents. During the comment and reply phases (24 September to 13 November) thirteen responses were made (with two further on the 26 November). On 12th December ICANN posted a report that summarized these responses.

All of this material can be found linked from:

http://www.icann.org/en/news/public-comment/whois-pp-abuse-study-24sep13-en.htm

Many of the comments were favourable, and very much appreciated. Other comments were directed at ICANN, making suggestions for future activity, and we do not consider those here. This Appendix is included in this final report to set out our responses to the criticisms of the draft report so that people can put those criticisms into context.

We now discuss the particular topics raised.

TOPIC A: Failure to study cybersquatting

The choice of what was to be studied was agreed during the negotiation phase of the contract prior to the endorsement of the NPL proposal by the GNSO Council. However, WP9 (UDRP disputes) does cover some aspects of cybersquatting.

Particular attention has been drawn to our suggestion that typosquatting is more prevalent than cybersquatting. In this study we looked at over 27,000 typosquatting domains in .com. The 2010 Moore & Edelman paper[29] estimated that at least 938,000 typosquatting domains existed at that time. At the same time they gave a figure of 45,000 UDRP actions and some actions under the US ACPA law which involved groups of a few hundred or perhaps a few thousand domains. Our figures indicate a reduction in typosquatting since the study was undertaken, but equally the number of UDRP actions over the period we considered was not all that high either. Naturally, publication of further statistics on typosquatting and cybersquatting would be valuable in assessing their relative prevalence.

TOPIC B: Failure to study 'software piracy'

The choice of what was to be studied was agreed during the negotiation phase of the contract prior to the endorsement of the NPL proposal by the GNSO Council.

We did suggest that WP7 (SURBL) would cover some aspects of this topic because we believed that spam emails remained an important vector for publicizing relevant websites. We would be the first to agree that the SURBL list is so wide-ranging that it is impossible to determine the contribution that 'software piracy' makes to our results.

TOPIC C: Failure to study 'media piracy'

The choice of what was to be studied was agreed during the negotiation phase of the contract prior to the endorsement of the NPL proposal by the GNSO Council.

[29] T. Moore & B. Edelmann: *Measuring the Perpetrators and Funders of Typosquatting*. 14th International Conference on Financial Cryptography and Data Security, LNCS 6052, Springer, pp.175–191, 2010. http://lyle.smu.edu/~tylerm/fc10typo.pdf

The original ICANN document (18 May 2010) setting out the requirements for the study specifically set out that 'media piracy' would involve the study of "domain names used by servers that illegally share copyrighted movies and music". This underpinned our comments about bandwidth and how mitigation efforts would concentrate on hosting arrangements. The comments made in response to the draft report are to the effect that "the most egregious" examples were Torrent sites and there was also concern about sites that provided links to infringing material. These definitions of the topic are out with the original description.

TOPIC D: Justifications for omitting categories are sometimes too brief

The choice of what was to be studied was agreed during the negotiation phase of the contract prior to the endorsement of the NPL proposal by the GNSO Council, and they were clearly satisfied with our explanations.

For the final version of this report we have rewritten the introduction to Section 18 to try and make it much clearer that a key reason for omitting many topics was that we were, and remain, unaware of any source of data which could be said to be relatively unbiased in coverage and global in scope. Without having a sufficiently large set of unbiased data it is simply not possible to study a topic in a scientific manner.

TOPIC E: Restriction to .biz, .com, .info, .net and .org

The decision to restrict analysis to just five top level domains was an ICANN decision, and makes this study's results consistent with those of other Whois studies. The report documents which other TLDs were encountered in each work package – in almost all cases .com completely dominated every other TLD.

TOPIC F: Lack of data about the percentages of usage of privacy or proxy services

The percentages of domain names used for lawful Internet activities that employ privacy or proxy services can be found in the tables in Section 12.

TOPIC G: Failure to consider alternative means of contacting domain registrants

The decision to use a phone number as a measure of the ability to make contact with a domain registrant was agreed during the negotiation phase of the contract, prior to the endorsement of the NPL proposal by the GNSO Council.

As we set out in Section 2.3, we did not think that using email to establish contact for the purposes of our survey would be an especially effective way of proceeding. We would have been able to measure deliverability but we were extremely pessimistic that a high proportion of registrants would respond to our question. With the phone calls we found that only a handful of people who answered the phone were entirely unhelpful.

A parallel study of Whois Misuse[30] by Nektarios Leontiadis and Nicolas Christin (collaborators on this study) achieved a 25% response rate to email invitations sent to a panel of law enforcement and security professionals asking them to take a survey on a topic that they might be expected to be highly motivated to give their opinions on.

However, the response rate in this parallel study for an emailed survey request sent to randomly chosen domain name registrants was 3.6%. We see no reason to believe that if we had used email in this study that the response rate for any of the registrants of legitimate domains would have been substantially higher than 3.6% and it is very unlikely indeed that

[30] http://www.icann.org/en/news/public-comment/whois-misuse-27nov13-en.htm

we would have been able to draw any statistically robust conclusions from this part of our investigation.

So we entirely agree with the comment that using email to contact domain registrants would have produced different results – but we believe that the large number of non-responsive registrants would have meant that these results would have been of no statistical value.

The discussions about the Registrar Accreditation Agreement (RAA) which was drawn to our attention post-dated our project proposal (which was originally submitted in 2010). Although we fully accept that using email when validating the correctness of Whois information at domain registration is less-invasive and less-sensitive than making a phone call, we cannot agree that any comparison with that process is especially relevant to our study.

It was also suggested that we should have attempted to contact registrants who used privacy or proxy services. This is of course impossible to do by telephone – it is only possible to reach these registrants by having a message relayed by the privacy or proxy service provider. Although some services provide unique email addresses for this purpose, many services do not do this.

ICANN have been considering a proposal to investigate how well registrants can be reached by relaying messages via privacy or proxy services. The Interisle Consulting Group was tasked to determine the feasibility of such a study and they reported on this in April 2012.[31]

TOPIC H: Scope of study

It was pointed out to us that the study goes beyond the scope and mandate for the study as set out in the 18th May 2010 ICANN document. This is entirely correct. Prior to award of contract, NPL recommended that the study might address further issues beyond those initially proposed, as we believed that this could add further value to the ICANN community and make a better use of our time and efforts. On 28 April 2011 the GNSO Council passed a resolution agreeing to this, which incorporated amendments put forward by the Registries Stakeholder Group, and awarded the contract to NPL.[32]

TOPIC I: Selection of groups for WP6

A comment was made that the selection of lawful and harmless activities in WP6 was "problematic" and excluded groups whose results "would have been more generalizable". The examples of omitted groups that were mentioned were "human rights organizations, minority rights organizations, religious organizations, political groups, as well as activist groups (political and others)".

We make it quite clear in Section 2.3 that our selection of "legal and harmless" activities in WP6 (and because of the way we can segment the URLs, in WP1 as well) was not intended to be generalizable to all domain names, explicitly saying "It is important to understand that the selection we have made is not necessarily representative of the overall usage of domain names for lawful and harmless reasons".

As we explain at the beginning of Section 12, our choice of lawful and harmless activities was intended to approximately mirror the criminal and harmful sites studied in some of the other work packages. However, our main aim was to study sufficient activities to be able to

[31] Interisle Consulting Group: WHOIS Proxy/Privacy Reveal & Relay Feasibility Survey, August 2012. http://gnso.icann.org/issues/whois/whois-pp-survey-final-report-22aug12-en.pdf

[32] https://community.icann.org/display/gnsocouncilmeetings/Motions+28+April+2011

demonstrate that there were wide variations in the usage of the use of privacy and proxy services by different types of lawful and harmless activity. From just the six types of activity we chose, we clearly showed this was the case, so there was no necessity to underline this point by extending the study to cover even more types of domain registrant.

TOPIC J: Validity of the second hypothesis

One of the responses said that our second hypothesis is "invalid". We don't understand what invalid means in this context – it may or may not be true (that's how one assesses a hypothesis) and our results show that it is only partially true.

This response went on say that "its results might have still been significantly different had the sampling of the study group, particularly that in WP6, been broadened to include lawful activities in the human rights and minority speech and activity area outlined in the paragraph above."

We found a wide variation in the use of privacy and proxy services by lawful and harmless activities – with some being significantly greater users, and some showing significantly less use. Adding another study group, whatever its level of privacy or proxy service usage might be, would not affect this finding.