

A Study of Whois Privacy and Proxy Service Abuse

Richard Clayton

Computer Laboratory, University of Cambridge,
Cambridge, CB3 0FD, UK.
`richard.clayton@cl.cam.ac.uk`

Tony Mansfield

National Physical Laboratory, Hampton Road,
Teddington, Middx., TW11 0LW, UK.
`tony.mansfield@npl.co.uk`

Abstract

When Internet domain names are registered for malicious purposes the registrants are unlikely to want their identity to be published in the Whois system. We show that around 90% of such registrants completely fail to provide valid contact phone numbers. However, there are significant variations, dependent upon the type of wickedness, as to whether or not privacy and proxy services are used as a way of hiding contact details. We also show that there is substantial use of privacy and proxy services for domain names registered for several types of lawful and harmless activity, but once again there are wide variations in the percentages involved. Our study is by far the largest to look at Whois information, making our results robust and giving a sound basis for our suggestions as to why the differences we measure may be occurring.

1 Introduction

The domain Whois system is a distributed database containing the contact details of the companies and individuals who have registered Internet domain names. This database is accessed by clients making Whois protocol requests to Whois servers operated by domain registries and registrars [3]. A parallel system, which we do not consider in this paper, records the contact details of the entities that have been allocated blocks of IP address space.

The Whois system is public and available for anyone to query, albeit rate restrictions may be imposed. Since some domain name registrants are unwilling to make their contact details public a number of ‘privacy’ and ‘proxy’ services have sprung up.

For our study We used ICANN’s¹ definition of privacy and proxy services [10]. When a domain is registered using a privacy service the registrant’s name is made public but the rest of the contact details are generic, suitable only for making contact with the privacy service. When a domain is registered with a proxy service the domain registrant is the proxy service itself and all of the details are those of the proxy service.

Some privacy and proxy services automate the relaying of email traffic to the entity actually operating the domain by providing unique, but entirely opaque, email addresses for each of the domains they are handling, but whether or not they do this does not affect whether they are deemed to be privacy or proxy services.

The purchasers of domain names that choose to use privacy and proxy services do so for a wide variety of reasons. Some are fearful of receiving email spam to their contact addresses, some wish to conceal their identities lest radical opinions expressed in cyberspace have real world

¹Internet Corporation for Assigned Names and Numbers: <http://icann.org>

consequences and some are registering the domains for use in criminal enterprises and wish to avoid identifying themselves to law enforcement.

Domain registrants could provide false information about their identity, but ICANN requires registrars to impose contractual provisions so that domain registrations may be cancelled when incorrect Whois information is found and not promptly corrected. Naturally, inaccurate identity information can be (and doubtless is) provided to privacy or proxy services, but in this latter case the data is not subject to public scrutiny and so is far less likely to be detected.

For some time, ICANN has been concerned about Whois inaccuracy and the potential for abuse and they have sponsored several studies with a view to obtaining reliable evidence as to how the domain name Whois system actually operates. This paper discusses a specific study, performed by the present authors, which was initially proposed in 2010 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.

We saw it as 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 broadened the groups of domains that we would study 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.

From the first we felt that it was inadequate to consider just these two hypotheses. We considered it essential to determine whether other methods are used by malicious domain registrants to obscure their identities, because in the event of changes to privacy and proxy services, it is very likely that malicious registrants will turn to these alternatives. Accordingly, we also set out to determine how many of the domain names we examined had been registered with incorrect Whois information – specifically whether or not we could reach the domain registrant using a contact telephone number from the Whois data.

Our proposal for a modified study was accepted by ICANN in April 2012 and the work was completed by summer 2013 with a draft report appearing in September 2013 [2]. The final report, addressing the feedback from the ICANN community, will be published in Spring 2014.

The formal reports are available for those wishing to inspect all of the detailed results. This paper presents this material in outline, but differs from the ICANN documents in that it goes beyond reporting the measurements we made and attempts to interpret these results and explain why they occur.

In Section 2 we discuss our experimental method; in Section 3 we outline the groups of domain name registrants that we studied; and in Section 4 we summarise our results. In Section 5 we discuss the role of Whois in countering abuse and attempt to explain the results we obtained. In Section 6 we present related work, including other studies that ICANN has sponsored, and then in Section 7 we conclude.

2 Experimental method

The basic approach of our study 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

Our study mainly addresses the use of domain names that have been implicated in illegal or harmful activities but we also examined (particularly in WP6) some samples of lawful and harmless domain names to establish a point of reference. However, these samples are not necessarily representative of the overall usage of domain names for lawful and harmless reasons.

Collecting domain names

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

We collected and examined the Whois data for the domain names within the five most popular 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. In the occasional cases where no Whois information could be obtained, we ignored the domain.

Identifying privacy and proxy services

It was extremely straightforward to identify the overwhelming majority of cases where privacy and proxy services had been used. Indeed, in many cases the Whois information contained specific text to this effect, albeit many proxy services (from the strict ICANN definition) styled themselves as privacy services using the more general meaning of the word. Equally, it was almost invariably straightforward to identify the cases where privacy and proxy services were not being employed.

For the handful of cases where there was doubt we consulted a commercial service which told us the total number of domains registered by any particular entity – and when there were thousands of domains we concluded that this was a proxy service; where there were only a few dozen we concluded that this was neither a privacy nor a proxy service.

For completeness (and to ensure that it was possible to cross-compare between different ICANN sponsored studies), we also followed NORC’s methodology from their 2010 Whois study [10] and checked for the presence of their list of keywords within the registrant details.

The Whois data was processed using the `Net::WHOIS::Deft-Whois` Perl package developed specifically for this project.² This provided the counts we needed of domains using privacy or proxy services, along with data on the contact phone numbers that had been provided.

Finding contact telephone numbers

If the domain was not using a privacy or proxy service we determined whether the Whois record contained a telephone number for the domain registrant. We then checked if these numbers were ‘apparently valid’ – that is to say that it looked plausible that the number could be used to telephone the registrant.

Telephone numbers with fewer than six digits or where the digits (apart from a country code) were all zeroes or all nines were immediately 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 their phone number. Numbers that were clearly invalid for the particular country because of incorrect length or unallocated area code were excluded.

Telephone survey of domain registrants

We created a random sample of approximately 200 domains per work package that had ‘apparently valid’ contact numbers. We gave the details of these 1453 domains to a sub-contractor, IID,³ one of whose specialities is in getting malware and phishing sites cleaned up, for which task they regularly speak on the phone to website owners all over the world.

IID employees then attempted to call the registrants of the domains in our sample to have a short conversation with them, in their native language, to ask whether they acknowledged registering the domain. We used the registrants’ declared locations to determine their timezones and we took care to check whether they were a business (48% of calls) or an individual (52%) so as to choose an appropriate time for the call. When calls were not answered the number was retried three times, on different times and days, to best improve the chance of making contact.

We did not attempt to make multiple calls to the same phone number, no matter how many domains it was the contact number for. We assumed that the same result would be obtained, no matter which domain we were calling about. We also assumed that the people we reached told us the truth, which we believe in general they did. There were a handful of registrants of domains associated with criminal activity who refused to answer our survey question. This occurred so seldom that it did not materially affect our results or the conclusions that we draw from them.

It might perhaps be questioned why we chose to make phone calls in this study, effectively using the presence and correctness of contact numbers as a proxy for the overall veracity of the registrant’s details. We were confident 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.

²WHOIS Data Extracted From Templates: <http://deft-whois.org>

³<http://internetidentity.com>

3 Description of work packages

We will now describe the nine groups of domains that we considered (work packages WP1 to WP9). Note that WP6 (lawful and harmless) is split into six separate groups of domains, and that our detailed analysis allowed us to split WP1 (phishing) into three distinct groups.

3.1 WP1: Phishing

Phishing is the creation of fake websites in order to steal security credentials. The URLs of these websites are mainly circulated by email and instant messaging systems.

We used the new URLs received from five lists ('feeds') of phishing URLs (a brand owner, two website takedown companies and two collating services) during the week of 18–24 April 2012. This yielded 16 420 unique phishing URLs which utilised 5 015 distinct domain names. The five gTLDs being studied compromised 56.9% of the total number of domains, and accounted for 8 110 (49.4%) of the URLs.

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

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

These are legitimate businesses whose domain name appears in a phishing URL because their service (free webhosting, URL shortening etc.) has been used for criminal purposes. Although one might expect the majority of these sites to have valid contact information in their Whois, they will provide other mechanisms for making contact to report abuse.

- 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. Their decisions about what information to place into Whois predates their site being compromised.

- Malicious registrations (we call this subset 'WP1m' in subsequent discussion)

These domains have been specially registered for use in phishing attacks and it is entirely probable that the registrant does not wish to be identified.

Picking out the WP1t domains was straightforward. Some of the domains are household names, and there are published lists of domains that provide 'cloud' services, web hosting, URL shortening etc.

Distinguishing between the WP1c and WP1m categories is generally very simple indeed. Intruders may only obtain partial access to a compromised website – so they are constrained to add their phishing pages deep within the directory hierarchy and this makes their URLs extremely easy to distinguish by searching for the relevant URL patterns.

However, where the URL was fairly generic, a manual process was required. 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.

3.2 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.

We recorded the 717 URLs listed in the 28 day period (18 Aug 2012 to 14 Sep 2012), yielding 714 domain names. There were 651 domains (91.2%) within the five gTLDs being studied.

3.3 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 on an affiliate basis – for example, affiliates send email spam or post irrelevant blog comments. 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.

The domain names we investigated 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 831 domain names for this study of which 764 (92.1%) were within the five gTLDs being studied.

3.4 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 mis-key the URL and thereby accidentally visit the typosquatting domain.

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 3 200 website names using .com (which were five or more characters long). They found that 80% of the sites hosted pay-per-click advertisements, often advertising the correctly spelled domain and its competitors [9].

For this study, Tyler Moore provided a list of typosquatting domains that were current in mid-January 2013 for the .com domains within the Alexa top 3 500 most popular websites. There were 27 006 domains – the largest group that we considered.

3.5 WP5: Child sexual abuse image websites

Child sexual abuse image websites (sometimes described as child pornography sites) are universally reviled and 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)⁴ – who operate a reporting hotline for this type of material – kindly agreed to provide data for this study.

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. For this study, just as in the other work packages, we only considered domains registered within .biz / .com/ .info / .net / .org, which were 597 in all, 91.0% of the total.

We did not conduct a telephone survey for this category because of the nature of the activity, the delay since domain registration, and to avoid interfering with any law enforcement operations.

3.6 WP6: Lawful and harmless websites

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. To provide a similar diversity of types of site, the categories were chosen to approximately mirror the criminal and harmful sites studied in some of the other work packages. However, it should be kept 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.

WP6.1: Banks

To correspond with the websites analysed in WP1 (phishing), we examined banking websites. We selected the domains 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 2 020 domains.

The five gTLDs being studied cover 1 679 domains (83.1% of the total). However, the Yahoo directory is poorly curated so we had to manually check the websites and exclude the non-banks. At the end of this validation process just 1 405 domains remained.

WP6.2: Executive search consultants

We wished to analyse an analogue of the domains from WP2 (advanced fee fraud) that seek to dupe people into becoming ‘money mules’ to receive and forward stolen money. We did not find a well-curated list of job recruitment companies with a global reach, so we used the list of members of Association of Executive Search Consultants. This gave us 256 domain names of which 183 (71.6%) were within the five gTLDs we were studying.

⁴<http://www.iwf.org.uk>

WP6.3: Law firms

Some of the domains in WP2 (advanced fee fraud) were fake law firms, used for scams involving fake inheritances. We decided to study the domains used by the members of Lex Mundi, which says it is “the world’s leading network of independent law firms”.⁵ This gave us 143 domains, of which 112 (78.3%) were in the five relevant gTLDs.

WP6.4: Legal pharmacies

WP3 considered the domain names used by unlicensed pharmacies, so we considered the list maintained by LegitScript of online pharmacies that they considered to be “safe for US patients”.⁶ This contained 264 pharmacies using 255 different domain names and all but 4 of them were included in the study.

WP6.5: Adult websites

WP5 considered domains involved in the distribution of child sexual abuse images, so we also considered a range of websites providing ‘adult’ content, that is to say legal websites that provide erotic material which would not be suitable for viewing except by consenting adults. We examined the “Business and Economy > Shopping and Services > Sex > Adult Galleries” section of the Yahoo directory which, on 1 April 2013 contained 3 758 entries, which used 3 594 domains. Excluding invalid entries, we ended up with 3 336 domains to study.

WP6.6: Typosquatted domains

As we described above, in WP4 we examined typosquatting domains - domain names that are a small variation of the domain name of a legitimate website within the Alexa top 3500. For this part of the study we considered the domain names that had been typosquatted, that is the domains from the Alexa top 3500 for which variants had been registered.

There were 1 227 domain names in this category. These domains are, by definition, being used by extremely popular websites – and of course they were not registered in the furtherance of any criminal activity. Since WP4 only considered .com domains, all of these domains are in .com.

3.7 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.⁷ They provided us with a feed of their ‘multi-surbl-list’ which aggregates various more specialist feeds.

The 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.

We recorded the 28 306 domains listed by SURBL over the week 18–24 July 2012. Between them, the five gTLDs being studied cover 20 956 domains (74.0% of the total) but excluding sub-domains of legitimate domains and missing Whois data reduced this to 20 763.

⁵<http://www.lexmundi.com/lexmundi/default.asp>

⁶<http://www.legitscript.com/>

⁷<http://www.surbl.org/>

3.8 WP8: Domains associated with malware (StopBadware domains)

StopBadware is a non-profit anti-malware organization working to prevent, mitigate, and remediate ‘badware’ websites – websites that serve up viruses, spyware, scareware, and other badware.⁸ They provided us with a list of 41 878 URLs – one week’s worth of new badware URLs, for the period 5–12 December 2012. These URLs used 20 450 distinct domains and 117 IP addresses. Between them, the five gTLDs being studied cover 10 926 domains (53.4% of the total). Once again we excluded sub-domains of legitimate domains and missing Whois, reducing the count to 10 833.

Because of the very generic nature of the lists that SURBL and StopBadware collate there is undoubtedly some overlap between the WP7 and WP8 categories. The WP7 (SURBL) list will almost certainly contain some examples of domain names associated with phishing and all the other categories of criminal activity we have studied.

The datasets come from very different time periods, so it is not possible to be certain about the exact size of the overlap, but because of the very different focus and list collation mechanisms used we do not believe that the overlap between WP7 and WP8 is substantial, and we also believe that the other types of criminal activity we have used specialised datasets to study, are just a small subset of the full WP7 and WP8 lists.

3.9 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).⁹ At present there are four approved dispute resolution providers, and we constructed our samples to, as closely as practical given some variance in reporting practice, reflect all the cases that were decided in January 2013.

A number of these cases involved multiple domain names, which were believed to have a common registrant. In these cases we analyzed the first domain of the set which came from the five gTLDs that we were studying. If there was no such domain then we just considered the first domain that was listed in the decision document. With this methodology, the 354 cases yielded 323 domains (91.2%) to be studied.

Rather than our usual practice of fetching Whois data as soon as we identified the domains to study, we used a commercial provider of archived Whois data to determine what information was being served immediately prior to the commencement of UDRP proceedings.

We did not do a telephone survey for the domains in this category because we would mainly have been ringing ex-registrants, some time after they lost control of the domain, and this could have influenced their responses.

4 Results of the study

We now report our results, starting with the exact counts of the usage of privacy and proxy services. This is followed by results inferred from our phone survey of samples of registrants, where we explain our use of an ‘a priori impossible to contact’ metric.

⁸<https://www.stopbadware.org/>

⁹<http://www.icann.org/en/help/dndr/udrp>

4.1 Usage of privacy and proxy services

This table gives the counts of domains studied, and the number of privacy/proxy registrations:

| | Work package | Domains | Privacy | Proxy | Both | %age |
|-------|-----------------------------------|---------|---------|---------|-------|-------|
| WP6.4 | Legal pharmacies | 251 | 16 + | 6 = | 22 | 8.8% |
| WP6.3 | Law firms | 112 | 14 + | 1 = | 15 | 13.4% |
| WP1t | Phishing: third parties | 263 | 2 + | 34 = | 36 | 13.7% |
| WP6.6 | Typosquatted domains | 1 227 | 48 + | 187 = | 235 | 19.2% |
| WP8 | StopBadware domains | 10 719 | 150 + | 2033 = | 2183 | 20.4% |
| WP6.2 | Executive search consultants | 183 | 24 + | 17 = | 41 | 22.4% |
| WP1c | Phishing: compromised sites | 2 121 | 37 + | 487 = | 524 | 24.7% |
| WP6.1 | Banks | 1 405 | 352 + | 44 = | 396 | 28.2% |
| WP5 | Child sexual abuse image websites | 597 | 19 + | 157 = | 176 | 29.5% |
| WP1m | Phishing: malicious registration | 449 | 29 + | 111 = | 140 | 31.2% |
| WP9 | Domains subject to UDRP | 320 | 4 + | 123 = | 127 | 39.7% |
| WP7 | SURBL domains | 20 763 | 74 + | 9091 = | 9165 | 44.1% |
| WP6.5 | Adult websites | 3 336 | 118 + | 1356 = | 1474 | 44.2% |
| WP2 | Advanced Fee Fraud | 651 | 21 + | 282 = | 303 | 46.5% |
| WP4 | Typosquatting | 27 006 | 378 + | 12642 = | 13020 | 48.2% |
| WP3 | Unlicensed pharmacies | 764 | 11 + | 408 = | 419 | 54.8% |

Summarising this in terms of the nature of each group of domains:

| | Work package | Maliciously registered? | Usage of privacy/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 |

Clearly, maliciously registered domains generally have a higher usage of privacy and proxy services than is the case for lawful and harmless activities. However, this correlation is not universal in that banks are above average users of these services, as are adult websites. We will discuss possible reasons for this in Section 5.

We should of course comment upon the statistical significance of the numbers we have just presented. As a rule of thumb, when the usage of privacy and proxy services in pairs of categories

differ by more than 3% then a using a χ^2 test shows that this is statistically significant, except that some of the sample sizes (particularly in WP6) are too small for this to be the case. The exact position is shown in the following table:

| | %age | WP6.4 | WP6.3 | WP1t | WP6.6 | WP8 | WP6.2 | WP1c | WP6.1 | WP5 | WP1m | WP9 | WP7 | WP6.5 | WP2 | WP4 | WP3 |
|---------------------------------------|-------|-------|-------|------|-------|-----|-------|------|-------|-----|------|-----|-----|-------|-----|-----|-----|
| WP6.4 Legal pharmacies | 8.8% | # | | | | | | | | | | | | | | | |
| WP6.3 Law firms | 13.4% | # | # | | | | | | | | | | | | | | |
| WP1t Phishing: third parties | 13.7% | # | # | # | | | | | | | | | | | | | |
| WP6.6 Typosquatted domains | 19.2% | # | # | # | # | | | | | | | | | | | | |
| WP8 StopBadware domains | 20.4% | # | # | # | # | # | | | | | | | | | | | |
| WP6.2 Executive search consultants | 22.4% | # | # | # | # | # | # | | | | | | | | | | |
| WP1c Phishing: compromised sites | 24.7% | # | # | # | # | # | # | # | | | | | | | | | |
| WP6.1 Banks | 28.2% | # | # | # | # | # | # | # | # | | | | | | | | |
| WP5 Child sexual abuse image websites | 29.5% | # | # | # | # | # | # | # | # | # | | | | | | | |
| WP1m Phishing: malicious registration | 31.2% | # | # | # | # | # | # | # | # | # | # | | | | | | |
| WP9 Domains subject to UDRP | 39.7% | # | # | # | # | # | # | # | # | # | # | # | | | | | |
| WP7 SURBL domains | 44.1% | # | # | # | # | # | # | # | # | # | # | # | # | | | | |
| WP6.5 Adult websites | 44.2% | # | # | # | # | # | # | # | # | # | # | # | # | # | | | |
| WP2 Advanced Fee Fraud | 46.5% | # | # | # | # | # | # | # | # | # | # | # | # | # | # | | |
| WP4 Typosquatting | 48.2% | # | # | # | # | # | # | # | # | # | # | # | # | # | # | # | |
| WP3 Unlicensed pharmacies | 54.8% | # | # | # | # | # | # | # | # | # | # | # | # | # | # | # | # |

A white square for the comparison of two different categories (on the X and Y axes) means the difference between two samples IS significant at the 90% level and a grey square (with a # symbol) indicates a non-significant difference. To make the pattern more clear, a grey square (with a = symbol) is used for the diagonal where a sample is being compared with itself.

We can also consider the results of the recent, May 2013, NORC study [11]:

| | | |
|-------------------------------|-------|-------|
| privacy or proxy registration | 320 | 20.0% |
| normal registration | 1 280 | 80.0% |
| TOTAL | 1 600 | |

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 paper, $\pm 1.6\%$.

We find, again 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 sampled from all registered domains whereas in all our categories there is a website involved – but 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 considered. That said, Table A-1 of their report 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 by legal persons had a 15.1% usage of privacy or proxy services. We see the lowest usages of privacy and proxy services within WP6 (lawful and harmless), which is considering activities generally undertaken by legal persons. That is, our findings are not grossly inconsistent with NORC’s results.

4.2 Reaching registrants by phone

We have already noted that when we examined the Whois data we determined whether or not there was an ‘apparently valid’ phone number. However, when we actually made the phone calls to the randomly chosen samples of domain registrants we found that some of these calls failed to connect – we treat this as a failure, equivalent to there being no phone number at all.

When we reached the person recorded in the Whois as being the registrant of the domain, but they denied all knowledge of it, we concluded that they had been impersonated. We treat this as a failure. If we did reach the registrant, or someone who could speak for them, and they acknowledged registering the domain then this counted as a success.

This left a number of scenarios which were neither success nor failure, such as when the number rang but was not answered; when we reached a voicemail account; or when the registrant was not available to come to the phone. We cannot tell whether, if we had been more persistent, the result would have been success or failure.

This ‘neither success nor failure’ category, along with the wide variations in usage of privacy/proxy services made it hard to understand what our data meant. We therefore decided to measure the likelihood that, given a domain name from a particular category, we would know that it was pointless deciding to make a phone call to the registrant of that domain name. This ‘a priori impossible to call’ status occurs when either a privacy or proxy service is used OR there is no phone number in the Whois data OR the phone number will not connect OR the person we reach will deny registering the domain. The first two values are of course exactly measured, the other two are scaled up from the results of our calls to the randomly chosen samples.

The results we obtained are as follows:

| | | sample size | calls made | impossible to call registrant | error bounds | maliciously registered? |
|-------|----------------------------------|-------------|------------|-------------------------------|--------------|-------------------------|
| WP6.4 | Legal pharmacies | 251 | 41 | 24.2% | 1.2% | no |
| WP6.3 | Law firms | 112 | 40 | 33.6% | 0.7% | no |
| WP6.2 | Executive search consultants | 183 | 40 | 36.7% | 0.9% | no |
| WP6.1 | Banks | 1 405 | 40 | 44.6% | 0.9% | no |
| WP6.6 | Typosquatted domains | 1 227 | 50 | 47.1% | 2.8% | no |
| WP1t | Phishing: third parties | 263 | 40 | 49.6% | 4.3% | no |
| WP8 | StopBadware domains | 10 719 | 201 | 51.4% | 4.4% | some |
| WP6.5 | Adult websites | 3 336 | 201 | 55.1% | 0.8% | no |
| WP7 | SURBL domains | 20 763 | 40 | 58.5% | 0.8% | mostly |
| WP1c | Phishing: compromised sites | 2 121 | 61 | 61.7% | 5.3% | no |
| WP4 | Typosquatting | 27 006 | 200 | 67.7% | 3.9% | yes |
| WP2 | Advanced Fee Fraud | 651 | 199 | 88.9% | 2.4% | yes |
| WP3 | Unlicensed pharmacies | 764 | 212 | 91.8% | 1.2% | yes |
| WP1m | Phishing: malicious registration | 449 | 99 | 92.5% | 7.5% | yes |

Note that we made no calls in two of the work packages, WP5 (child sexual abuse websites) because of the nature of the domains and WP9 (UDRP disputes) because we’d have mainly been ringing ex-registrants.

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). 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 is considerable variation in whether or not privacy or proxy services are used.

5 The role of Whois and discussion of our results

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 will almost always be in a position to deal with the website, or at the very least, communicate with their customer using their own internal records of how to reach them. So the first step is very often to identify the hosting company and that is done by looking up IP addresses in the appropriate Regional Internet Registry (RIR) Whois system rather than consulting the domain name Whois system which we consider here.

However, where a website has been compromised, action can sometimes be 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 a useful source of information.

Some websites use so-called ‘fast-flux’ techniques, where the 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 disrupt 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 maliciously – which can expedite action being taken.

Whois can be helpful in gaining an understanding of the scope of some activities. Even when the data is entirely falsified it is quite difficult to create sets of details that don’t have patterns to them, and these patterns can be used to group domains together. Linking domains by Whois (or by patterns of hosting or DNS usage) is of particular value because occasionally, perhaps at the start of their career, criminals will use their own name for a domain registration and some have proved to be traceable even if thereafter they are always careful to use fake details and/or privacy and proxy services.

5.1 Results relating to criminal activity

The results for the groups of domains involved in criminal activity are straightforward – nine times out of ten the criminal has not provided a phone number that will reach them.

In the case of child sexual abuse image websites we did not do a telephone survey, however, the IWF told us that, in their experience, when contact details were provided for these domains they were almost always extracted from online ‘white pages’ websites and were invariably false.

Given that one group of criminals is completely unreachable we might wonder why we can ever make a successful phone call to the registrant of a malicious domain? This may be because of experimental error – we’ve just misclassified the domain (this may have occurred in WP1m, maliciously registered phishing domains); it may be that some people have naïvely registered domains on behalf of criminals (this may be the case for some domains in WP2, advanced fee

fraud), but much of the explanation may just be that criminals are taking the view that in their jurisdiction they are not breaking any laws, or have no risk of being apprehended, and so being identified is not a risk.

The usage of privacy and proxy services differs markedly between different types of criminal activity. The reason for this may be the use that is made of Whois data in tackling the various crimes. Clearly, when making a report about a website containing child sexual abuse images it is not necessary to comment upon the contents of the Whois record, so whether or not a privacy or proxy service is used or not will hardly be relevant.

However, drawing attention to invalid Whois information may tip the balance when complaining about a 419 scam, where the website may otherwise appear legitimate. Additionally, complaining about inaccurate (or downright false) Whois information has been regularly used as a way to tackle unlicensed pharmacy domains. For example, Knujon have been running just such a campaign for several years [1], and using privacy or proxy services sidesteps this tactic.

5.2 Results relating to lawful and harmless activity

Our results for several types of lawful and harmless activity are somewhat variable, but easily distinguished from the results for criminal activities.

Legal pharmacies (WP6.4) use privacy and proxy services 8.8% of the time but the websites hosting adult material used these services 44.2% of the time – somewhat more often than several types of criminal activity.

However, all these registrants were, at least to some extent, contactable by telephone. Our success rate was highest for law firms (WP6.3) at 33.4% and banks (WP6.1) at 29.0%, 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 indeterminate call attempts, which neither totally failed nor totally succeeded, had worked out for us then our success rate would have doubled.

We believe what is happening is that we see more accurate phone number information being provided by large companies than by micro-businesses or individuals (who make up the bulk of the registrants in WP1c, the websites compromised for phishing). However, some large companies are choosing to remove contact details from Whois data to try and ensure that contact is only made through standard channels. The person who set up the bank's domain name doesn't want to be called out of hours when the online banking service has a problem.

Evidence that this is the likely explanation can be seen by examining the usage of privacy services. These are usually only a small fraction of the usage of proxy services, with the exception of WP6.1, WP6.2, WP6.3 and WP6.4 (banks, executive search consultants, law firms and legal pharmacies). Many of these domains were registered with Network Solutions – which offers a range of services going beyond that of a mass market registrar. Most other registrars proposed the use of a proxy service to registrants who wished to conceal details about themselves, but as you register a domain Network Solutions will propose the use of a privacy service.

The lowest call success rate for a legal activity was in WP6.5 (adult websites) where only 5.7% of registrants were reached in our survey and 55.1% of the domain registrants were impossible to reach by phone. That means of course that in nearly 40% of cases we can neither be sure of reaching a registrant by phone, nor sure of this being impossible. We do not know whether this reluctance to be reachable reflects the nature of the business being conducted or whether it is once again a wish to only deal with customers via other channels.

5.3 Results for the intermediate categories

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 (phishing).

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 ‘a priori 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 these 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 the way in which a handful of registrants respond can substantially affect the results of our survey – and the error bounds reflect this uncertainty.¹⁰

We suspect that there are some ‘report inflation’ effects occurring in the SURBL data in that attempts have been made (using passive DNS [12] and other techniques) to identify all the domains that could be used to mount an attack rather than just the one that is currently in use. This will tend to find groups of domains that are controlled by the same person. A related issue occurred with the phishing data in WP1 but, because we had full URLs and not just hostnames, we were able to remove the relevant URLs from our dataset before commencing our analysis. That was not possible with the WP7 and WP8 data.

5.4 Results for typosquatting and UDRP domains

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.

We find 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 WP4 (we made no phone calls in WP9) we reached the domain registrant 10.6% of the time – 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 paper by Moore and Edelmann [9] found that 63% of typosquatting domains displaying Google adverts used just five identifiers, that is only a handful of people are responsible for a great deal of this activity.

Typosquatting is a civil rather than criminal matter, so it might be expected that domain registrants were less cautious about revealing their identity; and conversely that it mattered

¹⁰As we noted earlier, our random sampling methodology for the telephone survey avoided attempting to make more than one phone call to the same registrant. This makes the calculation of the error bounds slightly complex – footnote 23 of our ICANN report gives the details.

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.

6 Related work

In 2009 the Chicago based National Opinion Research Center (NORC) collected Whois data on 2 400 domains for an ICANN commissioned study. They closely examined 1 419 domains of which 351 (24.7% \pm 2.2%) were using privacy or proxy services [10]. Later, ICANN revisited this data and checked all 2 400 domains finding that 429 (17.8% \pm 2.0%) were using privacy or proxy services [5]. We understand this change resulted from a refinement of the classification methodology.

In 2013 a further NORC investigation of a new set of 1 600 domains found that 320 (20.0% \pm 1.6%) were registered using privacy or proxy services [11].

There has been limited other work on Whois data. Felegyhazi et al. investigated the use of Whois data for proactive blacklisting of newly registered domains, but they did not consider registrant details, only the date of registration, the registrar and the name servers used [4]. Hence their results will have been unaffected by whether or not a privacy or proxy service was used. Similarly, Ma et al. extracted some simple features from Whois for their classifier but once again nothing that would have been affected by the use of privacy or proxy services [8].

In another ICANN sponsored study, Leontiadis and Christin attempted to measure how much misuse there was of Whois data in practice [6]. It is the perception that such misuse will occur, and that domain registrants will receive unwanted email, surface mail and phone calls, that is often cited as an important reason for legitimate registrants to choose to use privacy and proxy services. Although the study estimated that 44% of domain registrants would be affected by misuse, the low response rates to their surveys make this figure rather speculative.

7 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.

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).

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 analyze the URLs and determine whether the domain was registered by a third party such as a hosting service provider, by a legitimate business (or individual) whose website had been compromised or by someone who intended it be used for malicious purposes.

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 have them consult 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 NORC had measured by random sampling of all domains). Nearly two thirds of these registrants are impossible to contact by phone, and we reached only a quarter of them for our survey.

This pattern of those engaged in criminal activity choosing not to be contactable was replicated for the other types of activity we studied although the percentage choosing to use privacy and proxy services varied very widely.

A small 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.

So we might add to our initial hypotheses the following statements that we believe to be 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.

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.

These findings show the constraints on changing the domain Whois system to significantly improve accuracy and usefulness and to prevent misuse. Abolishing privacy and proxy services would affect a substantial amount of lawful activity, while criminals currently using these services might be expected to adopt the methods of their peers and instead provide incomplete and inaccurate data. Insisting that domain registration data was always complete and accurate would mean a great many lawful registrations would need to be updated.

Acknowledgements

Richard Clayton collaborated with NPL under EPSRC Grant EP/H018298/1, “Internet Security”. He is currently funded by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHSS&T/CSD) Broad Agency Announcement 11.02, the Government of Australia and SPAWAR Systems Center Pacific via contract number N66001-13-C-0131. This paper represents the position of the authors and not that of the aforementioned agencies.

The work would not have been possible without the invaluable assistance of Isaac Bright, Nicolas Christin, David Hindley, Fred Langford, Nektarios Leontiadis, 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.

References

- [1] G. Bruen: Evaluation of Internet Corporation of Assigned Names and Numbers (ICANN) compliance processes. Version 2, 2013. <http://www.icann.org/en/news/correspondence/bruen-to-chehade-22apr13-en.pdf>
- [2] R. Clayton and others: Whois Privacy and Proxy Service Abuse Draft Report. ICANN, 2013. <http://gnso.icann.org/en/issues/whois/pp-abuse-study-20sep13-en.pdf>
- [3] L. Daigle: WHOIS Protocol Specification. RFC 3912, IETF, 2004.
- [4] M. Felegyhazi, C. Kreibich and V. Paxson: On the Potential of Proactive Domain Blacklisting. Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more. USENIX Association, 2010.
- [5] Internet Corporation for Assigned Names and Numbers: ICANN Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the top 5 gTLDs. ICANN, 2010. <https://www.icann.org/en/resources/compliance/reports/privacy-proxy-registration-services-study-14sep10-en.pdf>
- [6] N. Leontiadis and N. Christin: WHOIS Misuse Study, draft report for public comment. ICANN, 2013. <http://whois.icann.org/sites/default/files/files/cmumisuse-study-26nov13-en.pdf>
- [7] N. Leontiadis, T. Moore, and N. Christin: Measuring and Analyzing Search-Redirection Attacks in the Illicit Online Prescription Drug Trade. In USENIX Security Symposium. 2011.

- [8] J. Ma, L. Saul, S. Savage and G.M. Voelker: Beyond blacklists: learning to detect malicious web sites from suspicious URLs. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.
- [9] T. Moore and B. Edelman: Measuring the perpetrators and funders of typosquatting. In: Financial Cryptography and Data Security, pp. 175–191. Springer, 2010.
- [10] NORC at the University of Chicago: Draft Report for the Study of the Accuracy of WHOIS Registrant Contact Information. ICANN, 2010. <https://www.icann.org/en/resources/compliance/reports/whois-accuracy-study-17jan10-en.pdf>
- [11] NORC at the University of Chicago: Whois Registrant Identification Study Project Summary Report. ICANN, 2013. <http://gnso.icann.org//en/issues/whois/registrant-identification-summary-23may13-en.pdf>
- [12] F. Weimer: Passive DNS replication. FIRST. 2005.