Empirically Measuring WHOIS Misuse*

Nektarios Leontiadis and Nicolas Christin

Carnegie Mellon University
leontiadis@cmu.edu, nicolasc@cmu.edu

Abstract. WHOIS is a publicly-accessible online directory used to map domain names to the contact information of the people who registered them (registrants). Regrettably, registrants have anecdotally complained about their WHOIS information being misused, e.g., for spam, while there is also concrete evidence that maliciously registered domains often map to bogus or protected information. All of this has brought into question whether WHOIS is still needed. In this study, we empirically assess which factors, if any, lead to a measurable degree of misuse of WHOIS data. We register 400 domains spread over the five most popular global top level domains (gTLD), using unique artificial registrant identities linked to email addresses, postal addresses, and phone numbers under our control. We collect, over six months, instances of misuse targeting our artificial registrants, revealing quantitative insights on both the extent and the factors (gTLD, domain type, presence of anti-harvesting mechanisms) that appear to have statistically-significant impact on WHOIS misuse.

Key words: WHOIS, misuse, security, privacy

1 Introduction

WHOIS is an online directory that primarily allows anyone to map domain names to the registrants' contact information. Based on their operational agreement with ICANN [2], all global Top Level Domain (gTLD) *registrars* (entities that process individual domain name registration requests) are required to collect this information during domain registration, and subsequently publish it into the WHOIS directory; how it is published depends on the specific *registry* used (i.e., entities responsible for maintaining an authoritative list of domain names registered in each gTLD). While the original purpose of WHOIS was to provide the necessary information to contact a registrant for legitimate purposes (e.g. abuse notifications, or other operational reasons), there has been increasing anecdotal evidence of misuse of the data made publicly available through the WHOIS service. For instance, some registrants[1] have reported that third-parties used their publicly available WHOIS information to register domains similar to the reporting registrants', using contact details identical to the legitimate registrants'. The domains registered with the fraudulently acquired registrant information were subsequently used to impersonate the owners of the original domains.

* This paper is derived from a study we originally conducted for ICANN [1].
[1] http://www.eweek.com/c/a/Security/Whois-Abuse -Still-Out-of-Control

While such examples indicate that legitimate registrants may suffer from misuse of their WHOIS data, registrants of malicious domains often use bogus information, or privacy or proxy registration services to mask their identities [3].

This sad state of affairs brings into question whether the existence of the WHOIS service is even needed in its current form. One suggestion is to promote the use of a structured channel for WHOIS information exchange, capable of authenticated access, using already available web technologies [4, 5, 6]. An alternate avenue is to completely abandon WHOIS, in favor of a new Registration Data Service. This service would allow access to verified WHOIS-like information only to a set of authenticated users, and for a specific set of *permissible purposes* [7].

The present paper attempts to illuminate this policy discussion by empirically characterizing the extent to which WHOIS misuse occurs, and which factors are statistically correlated with WHOIS misuse incidents. This research responds to the decision of ICANN's Generic Names Supporting Organization (GNSO) to pursue WHOIS studies [8] to scientifically determine if there is substantial WHOIS misuse warranting further action from ICANN.

We generalize previous work [9, 10] with a much more comprehensive study using 400 domains across the five largest global top level domains (*.COM*, *.NET*, *.ORG*, *.INFO* and *.BIZ*) which, in aggregate, are home to more than 127 million domains [11]. In addition, we not only look at email spam but also at other forms of misuse (e.g., of phone numbers or postal addresses).

We validate the hypothesis that public access to WHOIS leads to a measurable degree of misuse, identify the major types of misuse, and, through regression analysis, discover factors that have a statistically-significant impact on the occurrence of misuse.

The remainder of this paper is organized as follows. In Section 2 we provide an overview of the related work. We discuss our methodology in Sections 3 and 4. We present a breakdown of the measured misuse in Section 5, and the deployed WHOIS anti-harvesting countermeasures in Section 6. We perform a regression analysis of the characteristics affecting the misuse in Section 7, note the limitations of our work in Section 8, and conclude in Section 9.

2 Related Work

Elliot in [12] provides an extensive overview of issues related to WHOIS. Researchers use WHOIS to study the characteristics of various online criminal activities, like click fraud [13, 14] and botnets [15], and have been able to gain key insights on malicious web infrastructures [16, 17]. From an operational perspective, the Federal Bureau of Investigation (FBI) has noted the importance of WHOIS in identifying criminals, but the presence of significant inaccuracies hinder such efforts [18]. Moreover, online criminals often use privacy or proxy registration services to register malicious domains, complicating further their identification through WHOIS [3].

ICANN has acknowledged the issue of inaccurate information in WHOIS [19], and has funded research towards measuring the extent of the problem [20]. ICANN's GNSO, which is responsible for developing WHOIS-related policies, identified in [9]

Table 1: **Number of domains under each of the five global Top Level Domains within scope in March 2011 [11].**

gTLD	.COM	.NET	.ORG	.INFO	.BIZ	**Total**
# of domains	95,185,529	14,078,829	9,021,350	7,486,088	2,127,857	127,694,306
Proportion in population	75.54%	11.03%	7.06%	5.86%	1.67%	100%

the possibility of misuse of WHOIS for phishing and identity theft, among others. Nevertheless, ICANN has been criticized [12,21] for its inability to enforce related policies.

A separate three-month measurement study from ICANN's Security and Stability Advisory Committee (SSAC) [10] examined the potential of misuse of email addresses posted exclusively in WHOIS. The authors registered a set of domain names composed as random strings, and monitored the electronic mailboxes appearing in the domains' WHOIS records for spam emails, finding WHOIS to be a contributing factor to received spam. Our work adopts a similar but more systematic methodology, to measure a broader range of misuse types and gTLDs, examining five categories of domain names, over a period of six months.

3 Methodology

To whittle down the number of possible design parameters for our measurement experiment, we first conducted a pilot survey of domain registrants to collect experiences of WHOIS misuse. We then used the results from this survey to design our measurement experiment.

3.1 Constructing a Microcosm Sample

In November of 2011 we received from ICANN, per our request, a sample set of 6,000 domains, collected randomly from gTLD zone files with equal probability of selection. Of those 6,000 domains, 83 were not within the five gTLDs we study, and were discarded. Additionally, ICANN provided the WHOIS records associated with 98.7% (5,921) of the domains, obtained over a period of 18 hours on the day following the generation of the domain sample.

Out of these nearly 6,000 domains, we created a proportional probability microcosm of 2,905 domains representative of the population of 127 million domains, using the proportions in Table 1. In deciding the size of the microcosm we use as a baseline the 2,400 domains used in previous work [20], and factor in the evolution in domain population from 2009 to 2011.

Finally, we randomly sampled the domain microcosm to building a representative sample of $D = 1,619$ domains from 89 countries. (Country information is available though WHOIS.)

3.2 Pilot Registrant Survey

We use the domains' WHOIS information to identify and survey the 1,619 registrants associated with domains in D, about their experiences on WHOIS misuse. Further details on the survey questions, methodology, and sample demographics are available in the companion technical report [1].

Despite providing incentives for response (participation in a random drawing to be eligible for prizes such as iPads or iPods) we only collected a total of 57 responses, representing 3.4% of contacted registrants. As a result, this survey could only be used to understand some general trends, but the data was too coarse to obtain detailed insights.

With the actual margin of error at 12.7%, 43.9% of registrants claim to have experienced some type of WHOIS misuse, indicating that the public availability of WHOIS data leads to a measurable degree of misuse. The registrants reported that email, postal, and phone spam were the major effects of misuse, with other types of misuse (e.g. identity theft) occurring at insignificant rates.

These observations are based on limited, self-reported data, and respondents may incorrectly attribute misuse to WHOIS. Nevertheless, the pilot survey tells us that accurately measuring WHOIS misuse requires to primarily look at the potential for spam, not limited to email spam, but also including phone and postal spam.

3.3 Experimental Measurements

We create a set of 400 domain names and register them at 16 registrars (25 domains per registrar) across the five gTLDs, with artificial registrant identities. Each artificial identity consists of (i) a full name (i.e. first and last name), (ii) an email address, (iii) a postal address, and (iv) a phone number.

All registrants' contact details are created solely for the purpose of this experiment, ensuring that they are only published in WHOIS. Through this approach, we eliminated confounding variables. From the moment we register each experimental domain, and the artificial identity details become public through WHOIS, we monitor all channels of communication associated with every registrant. We then classify all types of communication and measure the extent of illicit or harmful activity attributed to WHOIS misuse targeting these registrants.

Given the wide variety of registrars and the use of unique artificial identities, the registration process did not lend itself to automation and was primarily manual. We registered the experimental domains starting in the last week of June 2012, and completed the registrations within four weeks. We then monitored all incoming communications over a period of six months, until the last week of January 2013. All experimental domains were registered using commercial services offered by the 16 registrars; we did not use free solutions like DynDNS.

4 Experimental Domain Registrations

We associated the WHOIS records of each of the 400 domains with a unique registrant identity. Whenever the registration process required the inclusion of an organization

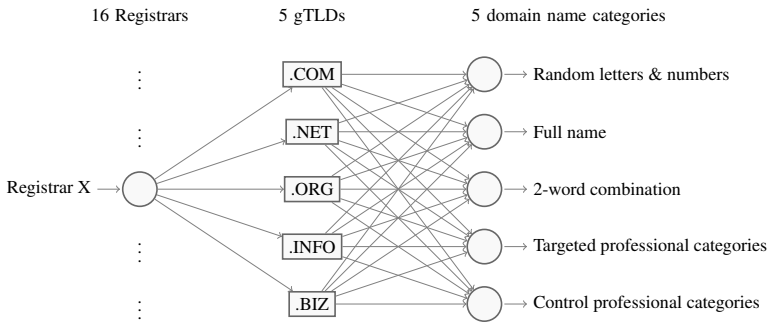


Fig. 1: **Graphical representation of the experimental domain name combinations we register with each of the 16 registrars.**

as part of the registrant information, we used the name of the domain's registrant. In addition, within each domain, we used the registrant's identity (i.e. name, postal/email address, and phone numbers) for all types of WHOIS contacts (i.e., registrant, technical, billing, and administrative contacts).

Figure 1 provides a graphical breakdown of the group of 25 domains we register per registrar. Every group contains five subgroups of domains, one for each of the five gTLDs. Finally, each subgroup contains a set of five domains, one for each type of domain name, as discussed later.

4.1 Registrar Selection

We selected the sixteen registrars used in our measurement study as follows. Using the WHOIS information of the 1,619 domains in D, we first identify the set R of 107 registrars used by domains in D. Some registrars only allow domain registration through "affiliates." In these cases we attempt to identify the affiliates used by domains in D, by examining the name server information in the WHOIS records.

We then sort the registrars (or affiliates, as the case may be) based on their popularity in the registrant sample. More formally, if $D_r \subset D$ is the set of domains in the registrant sample associated with registrar r, we define r's popularity as $S_r = |D_r|$. We sort the 107 registrars in descending order of S_r, and then select the 16 most popular registrars as the set of our *experimental registrars* that allow:

- The registration of domain names in all five gTLDs. This restriction allows us to perform comparative analysis of WHOIS misuse across the experimental registrars, and gTLDs.
- Individuals to register domains. Registrars providing domain registration services only to legal entities (e.g. companies) are excluded from consideration.
- The purchase of a single domain name, without requiring purchasing of other services for that domain (e.g. hosting).
- The purchase of domains without requiring any proof of identity. Given our intention to use artificial registrant identities, a failure to hide our identity could compromise the validity of our findings.

4.2 Experimental Domain Name Categories

We study the relationship between the category of a domain name, and WHOIS misuse. Specifically, we examine the following set of name categories:

1. Completely random domain names, composed by 5 to 20 random letters and numbers (e.g. unvdazzihevqnky1das7.biz).
2. Synthetic domain names, representing person full names (e.g. randall-bilbo.com).
3. Synthetic domain names composed by two randomly selected words from the English vocabulary (e.g. neatlimbed.net).
4. Synthetic Domain names intended to look like businesses within specific professional categories (e.g. hiphotels.biz).

To construct the last category, we identify professional categories usually targeted in cases of spear-phishing and spam, by consulting two sources. We primarily use the "Phishing Activity Trend" report, periodically published by the Anti-Phishing Working Group (APWG) [22]. We identify the professional categories mostly targeted by spam and phishing in the second quarter of 2010 with percentages of more than 4% in total. These categories are: (i) Financial services, (ii) payment services, (iii) gaming, (iv) auctions, and (v) social networking. We complement this list with the following professional categories appearing in the subject and sender portions of spam emails we had previously received: (i) medical services, (ii) medical equipment, (iii) hotels, (iv) traveling, and (v) delivery and shipping services.

In addition, we define a control set of professional categories that are not known to be explicitly targeted. We use the control set to measure the potential statistical significance of misuse associated with any of the previous categories. The three categories in the control set are : (i) technology, (ii) education, and (iii) weapons.

4.3 Registrant Identities

We create a set of 400 unique artificial registrant identities, one for each of the experimental domains. Our ultimate goal is to be able to associate every instance of misuse with a single domain, or a small set of domains.

A WHOIS record created during domain registration contains the following publicly available pieces of registrant information: (i) full name, (ii) postal address, (iii) phone number, and (iv) email address. In this section we provide the design details of each portion of the artificial registrant identities.

Registrant Name. The registrant's full name (i.e. first name-last name) serves as the unique association between an experimental domain and an artificial registrant identity. Therefore we need to ensure that every full name associated with each of the 400 experimental domains is unique within this context.

We create the set of 400 unique full names, indistinguishable from names of real persons, by assembling common first names (male and female) and last names with Latin characters.

Email Address. We create a unique email address for each experimental domain in the form ***contact***@*example.com*. We use this email address in the domain's WHOIS records, and we therefore call it *public email address*.

However, any email sent to a recipient other than *contact* (e.g. **foo**@example.com), is still collected for later analysis under a *catchall* account. We refer to these as *unpublished email addresses*, as we do not publish them anywhere, including WHOIS.

Mail exchange (MX) records are a type of DNS record pointing to the email server(s) responsible for handling incoming emails for a given domain name [23]. The MX records for our experimental domains all point to a single IP address functioning as a proxy server. The proxy server, in turn, aggregates and forwards all incoming SMTP requests to an email server under our control. The use of a proxy allows us to conceal where the "real" email server is located (i.e., at our university); our email server functions as a spam trap (i.e., any potential spam mitigation at the network- or host-level is explicitly disabled).

Postal Address. We examined the possibility of using a postal mail-forwarding service to register residential addresses around the world. Unfortunately, and, given the scale of this experiment, we were unable to identify a reasonably-priced and legal solution.

In most countries (the US included) such services often require proof of identification prior to opening a mailbox,[2] and limit the number of recipients that can receive mail at one mailbox. Moreover, we were hesitant to trust mail-forwarding services from privately owned service providers,[3] because the entities providing such services may themselves misuse the postal addresses, contaminating our measurements. For example, merely requesting a quote from one service provider, resulted in our emails being placed on marketing mailing lists without our explicit consent.

We eventually decided to use three Post Office (PO) boxes within the US; and, randomly assigned to each registrant identity one of these addresses. Traditionally, the address of a PO box with number *123* is of the following format: *PO Box 123, City, Zip code*. However, we utilize the US Postal Service's (USPS) *street addressing* service to camouflage our PO boxes as residential addresses. Street addressing enables the use of the post office's street address to reach a PO box located at the specific post office. Through this service, the PO box located at a post office with address *456 Imaginary avenue*, is addressable at *456 Imaginary avenue #123, City, Zip code*.

In addition, PO boxes are typically bound to the name of the person who registered them. However, each experimental domain is associated with a unique registrant name, even when sharing the same postal address, different than the owner of the PO box. We evaluated possible implications of this design in receiving postal mail to a PO box addressee not listed as the PO box owner. We originally acquired five PO boxes across two different US states, and sent one letter addressed to a random name to each of these PO boxes. We successfully received letters at three of the PO boxes indicating that mail addressed to any of the artificial registrant names would be delivered successfully. The test failed at the other two PO boxes—we got back our original test letters marked as undeliverable—making them unsuitable for the study.

[2] For example USPS form 1583: *Application for Delivery of Mail Through Agent* in the US.

[3] Also known as "virtual office" services.

Phone Number. Maintaining individual phone numbers for each of the 400 domains over a period of six months would be prohibitively expensive. Instead, we group the 400 domains into 80 sets of domains having the same gTLD and registrar, and we assign one phone number per such group. For example all .COM domains registered with GoDaddy share the same phone number.

We acquire 80 US-based phone numbers using Skype Manager[4] with area codes matching the physical locations of the three PO boxes. We further assign phone numbers to registrant identities with area codes matching their associated PO box locations.

5 Breaking Down the Measured Misuse

In this section we present a breakdown of the empirical data revealing WHOIS-attributed misuse. The types of misuse we identify fall within three categories: (1) *postal address misuse*, measured as postal spam, (2) *phone number misuse*, measured as voice mail spam, and (3) *email address misuse*, measured as email spam.

5.1 Postal Address Misuse

We monitor the contents of the three PO boxes biweekly, and categorize the collected mail either as *generic spam* or *targeted spam*. Generic spam is mail not associated with WHOIS misuse, while targeted spam can be directly attributed to the domain registration activity of the artificial registrant identities.

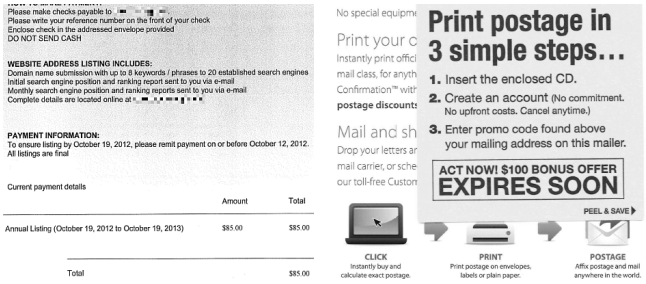
When postal mail does not explicitly mention the name of the recipient, we do not associate it with WHOIS misuse, and we classify it as generic spam. Common examples in this category are mail addressed to the "PO Box holder", or to an addressee not in the list of monitored identities.

In total, we collected 34 pieces of generic spam, with two out of the three PO boxes receiving the first kind of generic spam frequently. Additionally, we collected four instances of the second type of generic spam, received at a single PO box. A reasonable explanation for the latter is that previous owners of the PO box still had mail sent to that location.

Postal mail is placed in the targeted spam category when it is addressed to the name and postal address of one the of the artificial registrant identities. We observed targeted spam at a much lower scale compared to the generic spam, with a total of four instances.

Two instances of targeted postal spam, were sent to two different PO locations, but were identical in terms of (i) their sender, (ii) the advertised services, (iii) the date of collection from the PO boxes, and (iv) the posting date. The purpose of the letters, as shown in Figure 2a, was to sell domain advertising services. This advertising scheme works with the registrant issuing a one-time payment for $85 USD, in exchange for the submission of the registrant's domain to search engines in combination with search engine optimization (SEO) on the domains. The two experimental domains subjected to this postal misuse were registered using the same registrar, but under different registrant identities, and gTLDs.

[4] `http://www.skype.com/en/features/skype-manager/`



(a) Advertisement of search engine optimization services.

(b) Advertisement of postal and shipping services.

Fig. 2: **Targeted postal spam attributed to WHOIS misuse.**

The purpose of the third piece of targeted postal spam (Figure 2b) was to enroll the recipient in a membership program that provides postal and shipping services. Finally, the fourth piece of postal mail spam was received very close to the end of the experiment and offered a free product in exchange for signing up on a website.

Overall, the volume of targeted WHOIS postal spam is very low (10%), compared to the portion classified as generic spam (90%). However, this is possibly due to the small geographical diversity of the PO boxes.

5.2 Phone Number Misuse

We collected 674 voicemails throughout the experiment. We define the following five types of content indicative of their association (or lack thereof) to WHOIS misuse, and manually classify each voicemail into one of these five categories:
WHOIS-attributed spam. Unsolicited calls offering web-related services (e.g. website advertising), or mentioning an experimental domain name or artificial registrant name.
Possible spam. Unsolicited phone calls advertising services that cannot be associated with WHOIS misuse, given the previous criteria. (e.g. credit card enrollment based on random number calling)
Interactive spam. Special case of possible spam with a fixed recorded message saying "press one to accept".
Blank. Voice mails having no content, or with incomprehensible content.
Not spam. Accidental calls, usually associated with misdialing, or with a caller having wrong contact information (e.g. confirmation for dental appointment)

Two of these categories require further explanation. First, in the case of *possible spam*, we cannot tell if the caller harvested the number from WHOIS, or if it was obtained in some other way (e.g., exhaustive dialing of known families of phone numbers). We therefore take the conservative approach of placing such calls in a category separate from WHOIS-attributed spam. Second, calls marked as *interactive spam* did not contain enough content to allow for proper characterization of the messages. However, the large number of these calls—received several times a day, starting in the second month of the experiment—suggests a malicious intent.

Of the 674 voicemails, we classify 5.8% as WHOIS-attributed spam, 4.2% as possible spam, 38% as interactive spam, and 15% as not spam. Finally, we classify 36.9% of voicemails as blank due to their lack of intelligible content.

Of the 39 pieces of WHOIS-attributed spam, 77% (30) originated from a single company promoting website advertising services. This caller placed two phone calls in each of the numbers, one as an initial contact and one as a follow up. These calls targeted .BIZ domains registered with 5 registrars, .COM domains registered with 4 registrars, and .INFO domains registered with 6 Registrars. In total, the specific company contacted the registrants of domains registered with 11 out of the 16 registrars.

The remaining spam calls targeted .BIZ domains registered with 4 registrars, .COM domains registered with 4 registrars, and .INFO, .NET, and .ORG domains associated with 1 registrar each. In one case we observed a particularly elaborate attempt to acquire some of the registrant's personally identifiable information.

5.3 Email Address Misuse

We classify incoming email either as solicited or spam, using the definition of spam in [24]. In short, an email is classified as spam if (i) it is unsolicited, and (ii) the recipient has not provided an explicit consent to receive such email. For this experiment, this means that all incoming email is treated as spam, except when it originates from the associated registrars (e.g., for billing).

The contract between registrar and registrant, established upon domain registration, usually permits registrars to contact registrants for various reasons (e.g. account related, promotions, etc.). We identify such email by examining the headers of the emails received at the public addresses, and comparing the domain part of the sender's email address to the registrar's domain.

However, under the Registrar Accreditation Agreement (RAA) [2]), ICANN-accredited registrars are prohibited from allowing the use of registrant information for marketing, or otherwise unsolicited purposes. Nevertheless, we acknowledge the possibility that some registrars may share registrant information with third parties that may initiate such unsolicited communication. We do not distinguish between registrars that engage in such practices and those that do not, and we classify all communications originating from a party other than the registrar as spam.

Throughout the experiment, published email addresses received 7,609 unsolicited emails out of which 7,221 (95%) are classified as spam. Of the 400 experimental domains, 95% received unsolicited emails in their published addresses with 71% of those receiving spam email. Interestingly, 80% of spam emails targeted the 25 domains of a single registrar.

In an effort to explain this outlier, we reviewed the terms of domain registration for all 16 registrars. We discovered that four registrars (including the registrar that appears as an outlier) mention in their registrant agreements the possibility of use of WHOIS data for marketing purposes. Since this is only a hypothesis, we do not factor it into the regression analysis we propose later. It is, however, a plausible explanation for the outlier.

We classified all 1,872 emails received at the unpublished addresses as spam, targeting 15% of the experimental domains. Since the unpublished addresses are not shared

Table 2: **Breakdown of measured WHOIS-attributed misuse, broken down by gTLD and type of misuse.** Per the experimental design (Section 4), each gTLD group contains 80 domains.

	gTLD of affected experimental domains					
Type of misuse	.COM	.NET	.ORG	.INFO	.BIZ	**Total**
Postal address misuse	1 domain	1 domain	1 domain	1 domain	–	4 domains
Phone number misuse	5.0%	1.3%	1.3%	7.5%	10.0%	5.0%
Email address misuse	60.0%	65.0%	56.3%	77.5%	93.8%	70.5%

in any way, all emails received are unsolicited, and therefore counted as spam, including some that may have been the result of the spammers attempting some known account guessing techniques.

Two domains received a disproportionate amount of spam in their unpublished mailboxes. We ascribed this to the possibility that (i) these domains had been previously registered, and (ii) the previous domain owners are the targets of the observed spam activity. Historical WHOIS records confirm that both domains had been previously registered (12 years prior, and 5 years prior, respectively), which lends further credence to our hypothesis.

We examine the difference in proportions of email spam between published and unpublished addresses. Using the χ^2 test, we find that the difference is statistically significant considering the gTLD ($p < 0.05$), and the registrar ($p < 0.001$), but not the domain name category ($p > 0.05$).

Attempted Malware Delivery. We use VirusTotal [25] to detect malicious software received as email file attachments during the first 4 months of the experiment. In total, we analyze 496 emails containing attachments. Only 2% of emails with attachments (10 in total) targeted published email addresses, and they were all innocuous. The 15.6% of emails (76 in total) containing malware, targeted exclusively unpublished addresses, and VirusTotal classified them within 12 well-known malware families. As none of the infected attachments targeted any published email address, we do not observe any WHOIS-attributed malware delivery.

5.4 Overall Misuse per gTLD

In Table 2 we present the portion of domains affected by all three types of WHOIS misuse, broken down by gTLD and type of misuse. We find that the most prominent type of misuse is the one affecting the registrants' email addresses, followed by phone and postal misuse. Due to the small number of occurrences of postal misuse, we present the absolute value of affected domains. For both phone and email misuse, we present the misuse as the portion of affected domains, out of the 80 experimental domains per gTLD. Clearly, email misuse is common; phone misuse is also not negligible (especially for .BIZ domains).

The stated design limitations, especially the limited number of postal addresses we use, potentially affect the rates of misuse we measure. We nevertheless find that misuse of registrant information is measurable, and causally associated with the unrestricted availability of the data through WHOIS. We acknowledge though that this causal link is only valid based on the assumption that all ICANN-accredited registrars comply with the relevant RAA provisions (e.g., no resale of the registrant data for marketing purposes), as discussed in Section 5.3.

6 WHOIS Anti-Harvesting

WHOIS "anti-harvesting" techniques are a proposed solution deployed at certain registrars to prevent automatic collection of WHOIS information. We next present a set of measurements characterizing WHOIS anti-harvesting implemented at the 16 registrars and the three thick WHOIS registries.[5] Later on we use this information to examine the correlation between measures protecting WHOIS, and the occurrence of misuse.

More specifically, we test the rate-limiting availability on port 43, which is the well-known network port used for the reception of WHOIS queries, by issuing sets of 1,000 WHOIS requests per registrar and registry, and analyzing the responses. Each set of 1,000 requests repeatedly queries information for a single domain from the set of 400 experimental domains. We use different domain names across request sets. We select domains from the .COM and .NET pool when testing the registrars' defenses, and from the appropriate gTLD pool when testing thick WHOIS gTLD registries.

In addition, we examine the defenses of the remaining 89 registrars in the registrar sample. In this case we query domains found in the registrant sample instead of experimental domains. In three occasions, all domains associated with three out of the 89 registrars had expired at the time we ran this experiment. Therefore, we exclude these registrars from this analysis.

The analysis of WHOIS responses reveals the following methods of data protection:

Method 1: Limit number of requests, then block further requests.
Method 2: Limit number of requests, then provide only registrant name and offer instructions to access complete the WHOIS record through a web form.
Method 3: Delay WHOIS responses, using a variable delay period of a few seconds.
Method 4: No defense.

In Table 3 we present in aggregate form the distribution of registrars and registries using each one of the four defense methods. We find that one of the three registries does not use any protection mechanism, while the remaining two take a strict rate-limiting approach. For instance, one registry employs relatively strict measures by allowing only four queries though port 43 before applying a temporary blacklist.

Only 41.6% of the experimental registrars employ rate-limiting, allowing, on average, 83 queries, before blocking additional requests. Just two registrars in this group

[5] *Thick WHOIS* registries maintain a central database of all WHOIS information associated with registered domain names, and they respond directly to WHOIS queries with all available WHOIS information. From the five gTLDs under consideration, the three registries maintaining the .BIZ, .INFO, and .ORG zones are thick registries.

Table 3: **Methods for protecting WHOIS information at 104 registrars and three registries.**

Tested entities	**Total #**	**Type of WHOIS harvesting defense**			
		Method 1	Method 2	Method 3	Method 4
Thick WHOIS registries	3	2 (66.6%)	–	–	1 (33.3%)
Experimental registrars	16	7 (43.7%)	2 (12.5%)	1 (6.3%)	6 (37.5%)
Remaining registrars	89	37 (41.6%)	1 (1.1%)	3 (3.4%)	48 (53.9%)

provide information (as part of the WHOIS response message) on the duration of the block, which, in both cases, was 30 minutes. The remaining registrars either use a less strict approach (Method 2, 18.8%), or no protection at all (Method 4, 37.5%)

One registrar would not provide responses in a timely manner (method 3), causing our testing script to identify the behavior as a temporary blacklisting. It is unclear if this is an intended behavior to prevent automated queries, or if it was just a temporary glitch with the registrar.

The remaining 89 registrars (not in the experimental set) follow more or less the same pattern as our experimental set. The majority does not use any protection mechanism, and a relatively large minority uses Method 1.

7 Misuse Estimators

We finally examine the correlation of a set of parameters (i.e. estimators) with the measured phone and email misuse, attributed to WHOIS. These estimators are descriptive of the experimental domain names, and of the respective registrars and (thick) WHOIS registries. We do not examine postal address misuse, as the number of observed incidents in this case is very small and unlikely to yield any statistically-significant findings.

More specifically, we consider the following estimators:

- β_1 : Domain gTLD.
- β_2 : Price paid for domain name acquisition.
- β_3 : Registrar used for domain registration.
- β_4 : Existence of WHOIS anti-harvesting measures at the registrar level for .COM and .NET domains (thin WHOIS gTLDs), and at the registry level for .ORG, .INFO, and .BIZ domains (thick WHOIS gTLDs).
- β_5 : Domain name category.

We disentangle the effect of these estimators on the prevalence of WHOIS misuse through regression analysis. We use logistic regression [26], which is a generalized linear model [27] extending linear regression. This approach allows for the response variable to be modeled through a binomial distribution given that we examine WHOIS misuse as a binary response (i.e. either the domain is a victim of misuse or not).

In addition, using a generalized linear model instead of the ordinary linear regression allows for more relaxed assumptions on the requirement for normally distributed errors. In this analysis, we use the iteratively re-weighted least squares [28] method to fit

the independent variables into maximum likelihood estimates of the logistic regression parameters.

Our multivariate logistic regression model takes the following form:

$$logit(p_{\text{DomainEmailMisuse}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \quad (1)$$

$$logit(p_{\text{DomainPhoneMisuse}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (2)$$

Equation 2 does not consider β_5 as an estimator, since the experimental design does not permit the association between measured misuse and the composition of the domain name.

We considered the use of multinomial logistic regression (MLR) for the analysis of phone number misuse, given the five classes of voicemails we collected. Such regression models require a large sample size (i.e. observations of misuse in this case) to calculate statistically-significant correlations [29]. However, in the context of our experiment, the occurrence of voicemail misuse is too small to analyze with MLR.

Therefore, we reverted to using a basic logistic regression by transforming the multiple-response dependent variable into a dichotomous one. We did this by conservatively transforming observations of possible spam into observations of not spam. In addition, we did not consider the categories of interactive spam and blank, as they do not present meaningful outcomes.

All estimators, except β_2, represent categorical variables, and they are coded as such. Specifically, we code estimators β_1, β_3, and β_5 as 5-part, 16-part, and 5-part categorical variables respectively, using deviation coding. Deviation coding allows us to measure the statistical significance of the categorical variables' deviation from the overall mean, instead of deviations across categories.

We code WHOIS anti-harvesting (β_4) as a dichotomous categorical variable denoting the protection of domains by any anti-harvesting technique. While the 16 registrars, and 3 thick WHOIS registries employ a variety of such techniques (Section 6), the binary coding enables easier statistical interpretation.

7.1 Estimators of Email Misuse

In Table 4 we report the statistically-significant regression coefficients, and associated odds characterizing email misuse. Overall, we find that some gTLDs, the domain price, WHOIS anti-harvesting, and domain names representing person names are good estimators of email misuse.

Domain gTLD. The email misuse measured though the experimental domain names is correlated with all gTLDs but .INFO. Specifically, the misuse at .BIZ domains is 21 times higher than the overall mean, while domains registered under the .COM, .NET, and .ORG gTLDs experience less misuse.

Domain Price. The coefficient for β_2 means that each $1 increase in the price of an experimental domain corresponds to a 15% decrease in the odds of the registrants experiencing misuse of their email addresses. In other words, the more expensive the registered domain is, the lesser email address misuse the registrant experiences.

Table 4: **Statistically-significant regression coefficients affecting email address misuse (Equation 1).**

Estimator	coefficient	odds	Std. Err.	Significance
Domain gTLD (β_1)				
.COM	-1.214	0.296	0.327	$p < 0.001$
.NET	-0.829	0.436	0.324	$p = 0.01$
.ORG	-1.131	0.322	0.318	$p < 0.001$
.BIZ	3.049	21.094	0.566	$p < 0.001$
Domain price (β_2)	-0.166	0.846	1.376	$p < 0.001$
Lack of WHOIS anti-harvesting (β_4)	0.846	2.332	0.356	$p = 0.01$
Domain name composition (β_5)				
Person name	-0.638	0.528	0.308	$p = 0.04$

The reported correlation does not represent a correlation between domain prices and differentiation in the registrars' services. Even though we did not systematically record the add-on services the 16 registrars offer, we did not observe any considerable differentiation of services based on the domain price. Most importantly, we did not use any such service for any of the experimental domains we registered, even when such services were offered free of charge.

What this correlation may suggest is that higher domain prices may be associated with other protective mechanisms, like the use of blacklists to prevent known harvesters from unauthorized bulk access to WHOIS. However, such mechanisms are transparent to an outside observer, so we may only hypothesize on their existence and their effectiveness.

WHOIS Anti-Harvesting. The analysis shows that the existence of WHOIS anti-harvesting protection is statistically-significant in predicting the potential of email misuse. The possibility of experiencing email misuse without the existence of any anti-harvesting measure is 2.3 times higher than when such protection is in place.

Domain Name Category. We identify the category of domains denoting person names (e.g. randall-bilbo.com) as having negative correlation to misuse. In this case, the possibility of experiencing email address misuse is slightly lower than the overall mean.

This appears to be an important result. However, we point out that all the domain names in this category contain a hyphen (i.e. -), contrary to all other categories. Therefore, it is unclear whether the reported correlation is due to the domain name category itself, or due to the different name structure.

7.2 Estimators of Phone Number Misuse

The gTLD is the only variable with statistical significance in Equation 2. Table 5 presents the 3 gTLDs with a significant correlation to the measured WHOIS-attributed phone number misuse. Domains under the .BIZ and .INFO gTLDs correlate with 7.4

and 5.1 times higher misuse compared to the overall mean, respectively. On the other hand, .ORG domains correlate with lower misuse, being close to the mean.

There is no verifiable explanation as to why gTLD is the sole statistically-significant characteristic affecting this type of misuse. A possible conjecture is that domains usually registered under the .BIZ and .INFO gTLDs have features that make them better targets.

Table 5: **Statistically-significant regression coefficients in Equation 2.**

Estimator	**coefficient**	**odds**	**Std. Err.**	**Significance**
Domain gTLD (β_1)				
.INFO	1.634	5.124	0.554	$p = 0.003$
.ORG	-2.235	0.106	0.902	$p = 0.01$
.BIZ	2.000	7.393	0.661	$p = 0.002$

8 Limitations

Specific characteristics of the experimental design (e.g., cost limits) result in some limitations in the extent or type of insights we are able to provide.

In particular, we were not able to use postal addresses outside the United States, due to mail regulations requiring proof of residency, in most countries. In addition, "virtual office" solutions are prohibitively expensive at the scale of our experiment, and, as discussed earlier, could introduce potential confounding factors. Therefore, we were not able to gain major insights on how different regions, and countries other than the US are affected by WHOIS-attributed postal address misuse.

Similarly, we were not able able assign a unique phone number to each of the 400 artificial registrant identities. Instead, every phone number was reused by five (very similar) experimental domains. This design limits our ability to associate an incoming voice call with a single domain name, especially if the caller does not identify a domain name or a registrant name in the call. Nevertheless, we were able to associate every spam call with a specific [registrar, gTLD] pair.

9 Conclusion

We examined and validated through a set of experimental measurements the hypothesis that public access to WHOIS leads to a measurable degree of misuse in the context of five largest global Top Level Domains. We identified email spam, phone spam, and postal spam as the key types of WHOIS misuse. In addition, through our controlled measurements, we found that the occurrence of WHOIS misuse can be empirically predicted taking into account the cost of domain name acquisition, the domains' gTLDs, and whether registrars and registries employ WHOIS anti-harvesting mechanisms.

The last point is particularly important, as it evidences that anti-harvesting is, to date, an effective deterrent with a straightforward implementation. This can be explained by the economic incentives of the attacker: considering the type of misuse we observed, the value of WHOIS records appears rather marginal. As such, raising the bar for collecting this data ever so slightly might make it unprofitable to the attacker, which could in turn lead to a considerable decrease in the misuse, at relatively low cost to registrars, registries, and registrants.

Acknowledgments. This research was partially funded by ICANN. Input from the anonymous reviewers, from the participants to the WHOIS Misuse Webinar, and from several members of the ICANN community contributed significant improvements to this manuscript. We are also grateful for numerous discussions with Lisa Phifer, Liz Gasster, Barbara Roseman and Mary Wong, which led to considerable refinments in the design of the experiments. Finally, Tim Vidas provided invaluable support in setting up and maintaining our email infrastructure, Ashwini Rao assisted with some of the early testing scripts and documentation, and Patrick Tague helped us with testing some of the postal boxes.

References

1. Leontiadis, N., Christin, N.: WHOIS misuse study (March 2014) Available at http://whois.icann.org/sites/default/files/files/misuse-study-final-13mar14-en.pdf. Last accessed June 20, 2014.
2. ICANN: 2013 Registrar Accreditation Agreement. https://www.icann.org/resources/pages/approved-with-specs-2013-09-17-en. (2013) Last accessed June 20, 2014.
3. Clayton, R., Mansfield, T.: A study of Whois privacy and proxy service abuse. In: Proceedings (online) of the 13th Workshop on Economics of Information Security, State College, PA (June 2014)
4. Newton, A., Piscitello, D., Fiorelli, B., Sheng, S.: A restful web service for internet names and address directory services. USENIX; login (2011) 23–32
5. Sullivan, A., Kucherawy, M.S.: Revisiting WHOIS: Coming to REST. IEEE Internet Computing **16**(3) (2012)
6. Hollenbeck, S., Ranjbar, K., Servin, A., Newton, A., Kong, N., Sheng, S., Ellacott, B., Obispo, F., Arias, F.: Using HTTP for RESTful Whois services by Internet registries. (2012)
7. Expert Working Group on gTLD Directory Services: A next generation registration directory service. https://www.icann.org/en/groups/other/gtld-directory-services/initial-report-24jun13-en.pdf. (2013) Last accessed June 20, 2014.
8. ICANN. Generic Names Supporting Organization: Motion to pursue WHOIS studies. http://gnso.icann.org/en/council/resolutions#20100908-3. (2010) Last accessed June 20, 2014.
9. ICANN. Security and Stability Advisory Committee: Advisory on registrar impersonation phishing attacks. http://www.icann.org/en/committees/security/sac028.pdf. (2008) Last accessed June 20, 2014.
10. ICANN. Security and Stability Advisory Committee: Is the WHOIS service a source for email addresses for spammers? http://www.icann.org/en/committees/security/sac023.pdf. (2007) Last accessed June 20, 2014.

11. ICANN: gTLD–specific monthly registry reports. http://www.icann.org/sites/default/files/mrr/[gTLD]/[gTLD]-transactions-201102-en.csv. (February 2011) Last accessed June 20, 2014.
12. Elliott, K.: The who, what, where, when, and why of WHOIS: Privacy and accuracy concerns of the WHOIS database. SMU Sci. & Tech. L. Rev. **12** (2008) 141
13. Dave, V., Guha, S., Zhang, Y.: Measuring and fingerprinting click-spam in ad networks. In: Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication, ACM (2012) 175–186
14. Christin, N., Yanagihara, S., Kamataki, K.: Dissecting one click frauds. In: Proc. ACM CCS'10, Chicago, IL (October 2010) 15–26
15. Yarochkin, F., Kropotov, V., Huang, Y., Ni, G.K., Kuo, S.Y., Chen, I.Y.: Investigating dns traffic anomalies for malicious activities. In: Dependable Systems and Networks Workshop (DSN-W), 2013 43rd Annual IEEE/IFIP Conference on, IEEE (2013) 1–7
16. Li, Z., Alrwais, S., Xie, Y., Yu, F., Valley, M.S., Wang, X.: Finding the linchpins of the dark web: a study on topologically dedicated hosts on malicious web infrastructures. In: IEEE Symposium on Security and Privacy, IEEE (2013) 112–126
17. Leontiadis, N., Moore, T., Christin, N.: Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In: Proceedings of the 20th USENIX Security Symposium, San Francisco, CA (August 2011) 281–298
18. United States Congress. House Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property: Internet Domain Name Fraud: The U.S. Government's Role in Ensuring Public Access to Accurate WHOIS Data. H. hrg. U.S. Government Printing Office (September 2003)
19. WHOIS Task Force 3: Improving accuracy of collected data. http://gnso.icann.org/en/issues/whois-privacy/tor3.shtml (2003) Last accessed June 20, 2014.
20. NORC: Proposed design for a study of the accuracy of WHOIS registrant contact information (2009) Available online at https://www.icann.org/en/system/files/files/norc-whois-accuracy-study-design-04jun09-en.pdf. Last accessed June 20, 2014.
21. Watters, P.A., Herps, A., Layton, R., McCombie, S.: Icann or icant: Is whois an enabler of cybercrime? In: Cybercrime and Trustworthy Computing Workshop (CTC), 2013 Fourth, IEEE (2013) 44–49
22. Anti-Phishing Working Group: Phishing attack trends report - Q2 2010 (Janurary 2010)
23. Mockapetris, P.: Domain names – Implementation and specification (RFC 1035). Information Sciences Institute (1987)
24. The Spamhaus Project: The definition of spam. http://www.spamhaus.org/consumer/definition/. Last accessed June 20, 2014.
25. VirusTotal: Free online virus, malware and URL scanner. https://www.virustotal.com/. Last accessed June 20, 2014.
26. Hosmer Jr, D.W., Lemeshow, S.: Applied logistic regression. John Wiley & Sons (2004)
27. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. Journal of the Royal Statistical Society. Series A **135**(3) (1972) pp. 370–384
28. Del Pino, G.: The unifying role of iterative generalized least squares in statistical algorithms. Statistical Science **4**(4) (1989) pp. 394–403
29. Ye, F., Lord, D.: Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Analytic methods in accident research **1** (2014) 72–85