WHOIS REGISTRANT IDENTIFICATION STUDY

Project Summary Report

PRESENTED TO:
ICANN

PRESENTED BY:
NORC at the University of Chicago

MAY 23, 2013

NORC at the UNIVERSITY of CHICAGO

Executive Summary

The Internet Corporation for Assigned Names and Numbers (ICANN) has contracted NORC at the University of Chicago (NORC) to conduct the **WHOIS Registrant Identification Study.** This project is an exploratory examination of WHOIS data for a representative sample of gTLD domain names, using WHOIS Registrant Name and Registrant Organization values to classify the types of entities that register domains, including natural persons, legal persons, and privacy and proxy service providers.

NORC analyzed available web/FTP content associated with each sampled domain name to classify the types of entities that appear to be using those domains and the various types of activities associated with them. Additionally, we analyzed inter-relationships between these categories, seeking to provide a foundation for answering the following questions posed by the Government Advisory Committee (GAC):

- What is the percentage of registrants that are natural versus legal persons?
- What is percentage of domain name uses that are commercial versus non-commercial?
- What is the relative percentage of Privacy/Proxy use among legal persons?
- What is the relative percentage of Privacy/Proxy use among domains with commercial use?

This report is a summary of the activities undertaken to conduct and complete this project. Of primary interest are the interpretations of the statistical analysis. In particular, we focus on analyses related to the following three questions.

1) What differences exist in how domains are actually used for domains registered by natural persons versus domains registered by legal persons versus domains registered via proxy?

2) What differences exist between how domains users that are natural persons identify themselves, versus how domain users that are legal persons identify themselves?

3) What differences exist in how domains with any type of potentially commercial activity are identified in WHOIS versus domains with no observed potentially commercial activity?

In many cases, classification of the characteristics and activities were difficult to discern and often had to be coded as “unknown.” Unknowns that remained even after extensive investigation is an important study finding because they illustrate the degree of the difficulty experienced by those trying to use WHOIS data and Internet content to identify domain registrants and users.

Nevertheless, NORC has produced a coded set of data that is useful for its intended purpose—an exploratory study of registrant and domain user characteristics and the types of domain use activities.

With respect to answering the issues posed by the GAC:

- **Percentage of registrants that are natural versus legal persons**: Based on our analysis of the WHOIS records retrieved from a random sample of 1,600 domains from the top five gTLDs,
    - 39 percent (± 2.4 percent[1]) appear to be registered by legal persons
    - 33 percent (± 2.3 percent) appear to be registered by natural persons
    - 20 percent (± 2.0 percent) were registered using a privacy or proxy service.
    - We were unable to classify the remaining 8 percent (± 1.4 percent) using data available from WHOIS.
- **Percentage of domain name uses that are commercial versus non-commercial**: Per the GNSO Council's request, we attempted to categorize all observed monetary activities that in some countries might be legally considered "commercial activities," documenting a broad range of potentially commercial activities to enable multiple post-study interpretations that apply varied legal definitions. For example, because pay-per-click ads were found so frequently that they dominated this variable. We completed our analysis with and without pay-per-click ads to enable both interpretations of potentially commercial activity.

  Based on our analysis of web/FTP content retrieved from a random sample of 1,600 domains from the top five gTLDs,
    - When pay-per-click ads are included in the monetary activities that make up potentially commercial activity, 57 percent (± 2.4 percent) of all sampled domains were perceived to have potentially commercial activity.
    - When pay-per-click ads are <u>not</u> included in the monetary activities that make up potentially commercial activity, approximately 45 percent (± 2.4 percent) of all sampled domains were perceived to have potentially commercial activity.

[1] A 95 percent confidence level is used for all margin of error calculations, as well as statements concerning statistical significance.

- **Relative percentage of Privacy/Proxy use among legal persons**: Based on our analysis of the WHOIS records and the web/FTP content retrieved from a random sample of 1,600 domains from the top five gTLDs,
    - 15.1 percent (± 2.9 percent) of domains used by legal persons were registered using a privacy or proxy service.
- **Relative percentage of Privacy/Proxy use among domains with commercial use**: Based on our analysis of the WHOIS records and the web/FTP content retrieved from a random sample of 1,600 domains from the top five gTLDs,
    - 22.9 percent (± 2.7 percent) of domains with potentially commercial activity were registered using a privacy or proxy service.

Additional interesting findings related to the three focus questions for this study are:

1) Differences in how domains are used based on registrant type

- Domain names registered by legal persons were
    - More likely to be used by legal persons[2]—52.2 ± 3.9 percent, as compared to the entire sample's 36.6 percent.
    - Equally as likely to be used for some kind of potentially commercial activity —59.9 ± 3.9 percent, as compared to the entire sample's 56.6 percent.
    - Equally as likely to have WHOIS addresses in the U.S.—59.4 ± 3.9 percent, as compared to the entire sample's 56.9 percent.
    - More likely to be both registered and used by the same legal person—27.8 ± 3.5 percent, as compared to the entire sample's 16.8 percent.
    - More likely to be used by a for-profit entity—39.9 ± 3.8 percent, as compared to the entire sample's 25.6 percent.

[2] For entity and commercial activity classification definitions see the draft Working Definitions document prepared by the GNSO drafting team as of February 18th, 2009. http://gnso.icann.org/files/gnso/issues/whois/whois-working-definitions-study-terms-18feb09.pdf

- Domain names registered by natural persons were
  - More likely to be used by natural persons—10.4 ± 2.6 percent, as compared to the entire sample's 5.4 percent.
  - Equally as likely to be used for some kind of potentially commercial activity as the overall sample—55.4 ± 4.3 percent, as compared to the entire sample's 56.6 percent.
  - Less likely to have WHOIS addresses in the U.S.—46.0 ± 4.3 percent, as compared to the entire sample's 56.9 percent.
  - More likely to have undetermined domain user/registrant relationships—72.5 ± 3.9 percent, as compared to the entire sample's 54.8 percent.
  - More likely to be used by a non-business entity—11.8 ± 2.8 percent, as compared to the entire sample's 6.4 percent.
- Domain names registered using a Privacy/Proxy service were
  - More likely to be parked—30.7 ± 5.0 percent, as compared to the entire sample's 20.5 percent.
  - More likely to be used for some kind of potentially commercial activity—64.6 ± 5.2 percent, as compared to the entire sample's 56.6 percent.
  - More likely to be registered with a WHOIS address in the U.S.—74.3 ± 4.8 percent, as compared to the entire sample's 56.9 percent.
  - More likely to have a user/registrant relationship of a customer of a privacy/proxy service—92.8 ± 2.8 percent,[3] as compared to the entire sample's 20.4 percent.
  - More likely to be used by an entity with an unclear business structure—71.4 ± 4.9 percent, as compared to the entire sample's 65.7 percent.

[3] This relative percentage is not 100 percent because NORC's coding of this variable used the identity of the entity that presumably contracted a privacy service to register the domains. In such cases, the registered name holder's identity was not shielded, and we could determine the relationship with the domain user.

2) Differences in how kinds of domains users identify themselves based on domain registrant type

- Domain names used by legal persons were
    - More likely to be registered by legal persons—55.1 ± 4.0 percent, as compared to the entire sample's 38.6 percent.
    - More likely to be used for some kind of potentially commercial activity—79.8 ± 3.2 percent, as compared to the entire sample's 56.6 percent.
    - Equally likely to have WHOIS addresses in the U.S.—54.9 ± 4.0 percent, as compared to the entire sample's 56.9 percent.
    - More likely to also be registered by that legal person—35.5 ± 3.9 percent, as compared to the entire sample's 16.8 percent.
    - More likely to be used by for-profit businesses—60.7 ± 3.7 percent, as compared to the entire sample's 25.6 percent.
- Domain names used by natural persons were
    - More likely to be registered by natural persons—60.4 ± 10.2 percent, as compared to the entire sample's 32.8 percent.
    - Less likely to have potentially commercial activity—36.8 ± 10.1 percent, as compared to the entire sample's 56.6 percent.
    - Equally likely to have WHOIS addresses in the U.S.—49.9 ± 10.4 percent, as compared to the entire sample's 56.9 percent.
    - More likely to be registered by that natural person—69.7 ± 9.6 percent, as compared to the entire sample's 16.8 percent.
    - Never used by a business; this is by design—when coding apparent business structure, if the user was a natural person, then the business structure was coded as not a business.

3) Differences in domains with potentially commercial activity (pay-per-clicks ads included)

- Domain with detected potentially commercial activity were
    - More likely to have legal person users—51.5 ± 3.3 percent, as compared to the entire sample's 36.6 percent.
    - Less likely to have user/registrant relationships that cannot be determined—44.8 ± 3.2 percent, as compared to the entire sample's 54.8 percent.
    - Less likely to have an unclear business structure—55.2 ± 3.2 percent, as compared to the entire sample's 65.7 percent.
- For both Apparent Registrant Type and Registrant WHOIS Address County/Region of the World differences between the relative percentage among domains with potentially commercial activity and the entire sample's percentage are small. Thus, knowing that a domain has potentially commercial activity does not provide any additional insight as to the registrant type or the WHOIS address of the registrant.

1: Introduction and Purpose

The Internet Corporation for Assigned Names and Numbers (ICANN) has contracted NORC at the University of Chicago (NORC) to conduct the **WHOIS Registrant Identification Study.** This project is an exploratory examination of WHOIS data for a representative sample of top five gTLDs, using WHOIS Registrant Name and Registrant Organization values to classify the types of entities that register domains, including natural persons, legal persons, and privacy and proxy service providers. The underlying intent of the study is to seek a foundational understanding of the types of entities and kinds of activities observed in gTLDs. Accordingly, the categories of entities and activities were not predetermined in this study, but were generated as NORC examined active websites and their domain name's WHOIS data.

NORC analyzed available web/FTP content associated with each sampled domain name to classify the types of entities that appear to be using those domains and the various types of activities associated with them. Additionally, we analyzed inter-relationships between these categories, seeking to provide a foundation for answering the following questions posed by the Government Advisory Committee (GAC)[4]:

- What is the percentage of registrants that are natural versus legal persons?
- What is the percentage of domain name uses are commercial versus non-commercial?
- What is the relative percentage of Privacy/Proxy use among legal persons?
- What is the relative percentage of Privacy/Proxy use among domains with commercial use?

We further developed entity and commercial activity classifications to help the ICANN community better understand the wide variety of issues and the potential implications on policy. We used the draft *Working Definitions* document prepared by the GNSO drafting team as of February 18, 2009[5] and the *Revised Terms of Reference for WHOIS Registrant Identification Studies*[6] as a starting point for all entity and commercial activity classification. We also built upon the entity classification and Privacy/Proxy service identification methodologies developed in previous studies: *Study of the Accuracy of WHOIS Registrant*

[4] http://www.icann.org/en/news/correspondence/karlins-to-thrush-16apr08-en.pdf

[5] http://gnso.icann.org/files/gnso/issues/whois/whois-working-definitions-study-terms-18feb09.pdf

[6] http://gnso.icann.org/issues/whois/tor-whois-registrant-id-studies-20may11-en.pdf

Contact Information[7] and *ICANN's Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the Top 5 gTLDs.*[8]

This report is a summary of the activities undertaken to conduct and complete this project and the key findings which emerged from this study. Of primary interest are NORC's interpretations of the sampled data – that is, our statistical analysis. We have gathered a set of data that is useful for its intended purpose—an exploratory study of registrant and domain user characteristics and the types of domain use activities. In particular, we focus on analyses related to the following three questions.

1) What differences exist in how domains are actually used for domains registered by natural persons versus domains registered by legal persons versus domains registered via proxy?

2) What differences exist between how domains users that are natural persons identify themselves, versus how domain users that are legal persons identify themselves?

3) What differences exist in how domains with any type of potentially commercial activity are identified in WHOIS versus domains with no observed potentially commercial activity?

In many cases, classification of the characteristics and activities were difficult to discern and often had to be coded as "unknown." Therefore, an understanding of the methodology used to collect and code the data is needed to fully recognize the implications of this analysis.

[7] http://www.icann.org/en/news/public-comment/whois-accuracy-study-15feb10-en.htm

[8] http://www.icann.org/en/compliance/reports/privacy-proxy-registration-services-study-14sep10-en.pdf

This report has five primary sections and three appendices. The first section, **Introduction and Purpose**, is a summary of the purpose for the **WHOIS Registrant Identification Study**. The second section, **Methodology**, provides a summary of the sampling, data collection and data coding methods used to put together the analysis dataset. Summaries of the coded variables are also provided. The **Relationship Analysis Results** section is a summary of the analyses NORC conducted with the dataset to see if significant relationships (or associations) exist between variables. In particular, we highlight findings related to the GAC questions stated above. The fourth section, **Lessons Learned**, is a summary of the lessons learned in conducting this study. The knowledge gained related to the best practices for extracting domain content may be valuable for future WHOIS protocol and policy development, as well as for future ICANN studies. The final section, **Conclusions and Recommendations**, provides overall conclusions. **Appendix A: Exploratory Analysis Report** is a detailed review of all the analysis comparisons NORC considered. This report includes the results highlighted in this document, and provides additional views of the collected information that may be of interest to the ICANN community. **Appendix B: Variable Glossary** is a summary of the variables used in the analyses presented in this report. **Appendix C: Report Modifications** is a summary of the differences between the February 6, 2013 *Draft Project Summary Report*[9] and this final report.

[9] The draft report is available from ICANN at http://www.icann.org/en/news/public-comment/whois-regid-15feb13-en.htm.

2: Methodology

2.1. Sample Selection

By its nature, an exploratory study is generally not designed to answer specific research questions. Rather, the sample selection is designed to be large enough to explore interesting features apparent in the data, but statistically significant results are not paramount. For this study, we wanted to be able to provide statistically accurate information to answer the GAC questions stated in section 1 while providing a good exploratory dataset. With this proviso, we specified a sample size of 1,600 because it allows a proportional estimate's margin of error at the 95 percent confidence level to be no more than plus/minus 5 percent for any subgroups with 400 or more domains (25 percent of the 1,600 sampled domains).[10] Smaller subgroups can still reach this level of statistical accuracy, but not for proportions that are close to 0.5.

According to the *June 2011 Registry Operator Monthly Reports*,[11] approximately 98.5 percent of all gTLD domain names are registered in the five largest gTLDs: *.com, *.net, *.org, *.info, and *.biz. Therefore, with agreement from ICANN, we designed a sample of 1,600 domain names from the top five gTLDs. The gTLDs *.edu, *.mil, and *.gov were deemed out of scope for this study because they are not administered by ICANN.

A gTLD stratified sample was selected by ICANN staff according to NORC specifications. Due to the small size of the *.biz domain, a proportional sample among the top five gTLDs would only select approximately 26 *.biz domain names, which would likely not be enough to make useful comparisons across domain name extensions.[12] A proportional sample would select 95 domain names from the *.info gTLD. In order to provide some information about each of the five gTLDs, NORC specified that 100 selections be made from each of the *.info and *.biz gTLDs, with the remaining 1,400 selected proportionally among the top three gTLDs. This results in a slight under-sample of *.com, *.net, and *.org domains, and an oversample *.biz and *.info domains. Case weighting is used to account for this

[10] Two of the GAC questions are related to overall proportional estimates: natural person registrants versus legal person registrants, and commercial domain use. These proportional estimates will be based the overall sample of 1,600 domain, so the accuracy goals will be met. The other two GAC questions concern the use of Privacy/Proxy services among subgroups—legal person users and domains with commercial use. We did not know *a priori* that there would be at least 400 domains with legal person users or with commercial use. ICANN's *Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service* concluded that approximately 24 percent of domains in the top five gTLDs are likely registered using a privacy or proxy service. Thus, we expect to find close to 400 domains overall that use Privacy/Proxy services. In this sense, we felt that a sample of 1,600 domains would provide accurate estimates in answer to the GAC questions.

[11] http://www.icann.org/en/tlds/monthly-reports/

[12] Comparative analysis across domain name extensions is summarized in Appendix A, section E.

when analyzing the results. Exhibit 1 is a summary of the sample selection and related weighting scheme for the five gTLDs.

Exhibit 1: Sample Design and Weight Factors for the Registrant ID Study

gTLD	Global Proportion	Sample Size	Sample Proportion	Weight = Global/Sample Proportion	Sum of Weights = Sample Size *Weight
*.com	74.3%	1,128	70.5%	1.0534	1188.2
*.net	10.7%	165	10.3%	1.0412	171.8
*.org	7.2%	107	6.7%	1.0813	115.7
*.info	6.1%	100	6.3%	0.9830	98.3
*.biz	1.6%	100	6.3%	0.2600	26.0
TOTAL	**100.0%**	**1,600**	**100.0%**		**1,600.0**

NORC proceeded to collect information for the selected domains, which included WHOIS information, as well as HTTP/HTTPS/FTP (web/FTP) content associated with sampled domain names. [13]

2.2. Data Collection

NORC constructed an automated information-gathering tool (the NORC-BOT) to collect data from multiple sources for a given domain. Unlike a web spider, which generally only crawls over HTTP content, the NORC-BOT attempted to collect the following three content sources for each domain: WHOIS data (WHOIS), publically accessible HTTP/FTP files, and response codes from DNS BlackLists (DNSBL) for the given domain.

Due to the fact that the World Wide Web (www) is not a static environment it is possible that domain information could change at any given moment. Domain registration changes, WHOIS record updates, web content changes, etc. are all part of the dynamic nature of the Internet environment. Because of the fluid nature of this content, the amount of time that lapses between extractions introduces a potential for data stagnation error. As a result of the lack of information on how often the content sources update, it is impossible to measure the extent of the data stagnation error. However, reducing the amount of time between the content source extractions for a given domain should reduce the potential for this error to occur. Therefore, NORC attempted to simultaneously conduct the three content source extractions.

[13] Only the primary website hosted at each sampled domain was searched for content. This included the www.domain and ww2.domain for both HTTP and FTP sites. Domains may have been associated with other servers or uses, and no attempt was made to look for such content.

To strive towards simultaneous extraction, we created the multi-threaded application NORC-BOT. Its programming is written in the Python (version 2.7) language, a dynamic programming language facilitating rapid application development. Taking advantage of these benefits NORC was able to develop an application that distributes the tasks associated with extracting content sources across multiple threads. These threads ran in parallel thus providing simultaneous extraction of the content sources for a given domain. However, NORC-BOT obtained WHOIS information from the *WhoisAPI* service.[14] Initial runs of the NORC-BOT revealed that *WhoisAPI* did not always return WHOIS information for all the domains (see the **Lessons Learned** section of this report for more details). Therefore, ICANN and NORC decided that the most effective way to collect WHOIS data would be for ICANN to run its own WHOIS data extraction in parallel with the NORC-BOT run. All domains with no WHOIS information from the NORC-BOT would be able to be merged with the ICANN-extracted WHOIS dataset. The two WHOIS datasets were also compared to verify consistency.

2.3. Data Coding

After the domain content was extracted by the NORC-BOT on March 16, 2012, we undertook an effort to create an analysis dataset of coded variables. In general, the coded variables fall into three broad classes: 1) WHOIS variables, 2) Domain User variables, and 3) Domain Content variables. We provide an overview of each of these classes and provide a summary of some of the key analysis variables in each class. In many cases, classification of the characteristics and activities were difficult to discern and often had to be coded as "unknown." These unknown classifications illustrate the degree of the difficulty experienced by those trying to use WHOIS data and Internet content to identify domain users. The Lessons Learned section of this report provides a summary the difficulties NORC encounter with coding some of the variables.

2.3.1. WHOIS Variables

WHOIS variables in the final coded dataset are those associated with information collected by the NORC-BOT from the WHOIS via *WhoisAPI*, a web-based service that returns machine-parsable WHOIS fields for a domain through an HTTP request. This information was supplemented by WHOIS information extracted for each sampled domain by a process developed by ICANN staff. For cases in which the *WhoisAPI* service did not send back a reply to the NORC-BOT request, ICANN extracted WHOIS information was used.

[14] http://whoisxmlapi.com/

The extracted WHOIS information was cleaned and manually processed in order to correct parsing errors, overcome inconsistencies in WHOIS data, and establish the registrant name and organization and registrant country. Additional manual coding was done to determine the apparent registrant type and the registrar.

Apparent Registrant Type

Apparent registrant type was coded as to whether we could place the registrant into categories defined in ICANN's *Revised Terms of Reference for WHOIS Registrant Identification Studies*.[15] Using methodology described below, these categories were collapsed into the following four general Apparent Registrant Types:

- Registrant appears to be a Legal Person – domains with WHOIS data which appear to identify a legal person—a company, business, partnership, non-profit entity, trade association, etc.—as the Registrant (includes multiple domain holders, but not Privacy/Proxy service providers)
- Registrant appears to be a Natural Person – domains with WHOIS data which appear to identify a natural person—a real, living individual—as the Registrant
- Registrant appears to reference a Privacy/Proxy Service – domains with WHOIS data which appear to identify a Privacy/Proxy service
- Unclassified – domains which could not be classified using WHOIS data (includes data completely missing, patently false, or incomplete, and domains pending reactivation or deletion)

Initially, only WHOIS information and independent searches of public databases were considered in the classification. For example, we searched known lists of privacy and proxy providers—identified while conducting previous WHOIS related studies[16]—to place sample domain registrants into these categories, and reverse WHOIS[17] email counts were used to help determine multiple domain name holders. While investigating the domain user the coder may have gained insights on the registrant of the domain, such as situations where the domain user is the same as the registrant. Thus, additional information was used to correct initial categorizations or add granularity to the process. In the end, registrant type was based on

[15] http://gnso.icann.org/issues/whois/tor-whois-registrant-id-studies-20may11-en.pdf

[16] "Study of the Accuracy of WHOIS Registrant Contact Information" (http://www.icann.org/en/news/public-comment/whois-accuracy-study-15feb10-en.htm), and "Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service" (http://www.icann.org/en/compliance/reports/privacy-proxy-registration-services-study-14sep10-en.pdf)

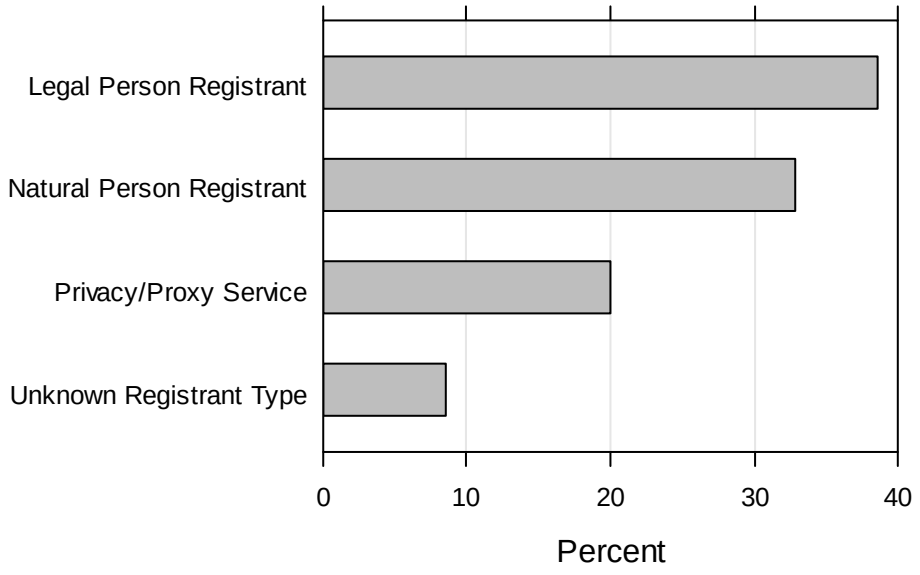
[17] A reverse WHOIS lookup allows one to search for a specific entity's name in WHOIS records. . NORC used the service at http://reversewhois.domaintools.com/. A uniquely identifying piece of information about any specific person or company (like their name, email address or phone number) is entered into the search engine. Reverse Whois provides a report of all the current domain names containing that information in the WHOIS records.

the evidence that we were able to discover during our investigation, that is, each domain registrant was coded based on what was apparent in the information we found in the dataset. No attempt was made to verify WHOIS accuracy or contact the identified registrant.

Exhibit 2 provides a summary of Apparent Registrant Type for the sampled domains, and provides information to answer the GAC question: ***What is the percentage of registrants that are natural versus legal persons?***

> In our random sample of 1,600 domains, 39 percent (± 2.4 percent[18]) appear to be registered by legal persons and 33 percent (± 2.3 percent) appear to be registered by natural persons—a statistically significant difference. Another 20 percent (± 2.0 percent) were registered using a privacy or proxy service. We were unable to classify the remaining 8 percent (± 1.4 percent) using data available from WHOIS.

Exhibit 2: Apparent Registrant Type for Sampled Domains



Further analysis of Apparent Registrant Type is summarized in section 3.1. We review relationships between Apparent Registrant Type and each of the following variables: Apparent Domain User Type, Potentially Commercial Activity, Apparent Business Structure, Registrant's WHOIS Address Country/Region of the World, and the Domain User Relationship with the Registrant. Additional analysis of Apparent Registrant Type can be found in Appendix A.

[18] A 95 percent confidence level is used for all margin of error calculations, as well as statements concerning statistical significance.

Privacy/Proxy Services

As shown in **Exhibit 2**, Privacy/Proxy services were used to register 20 percent of the sampled domain (320 domains out of 1,600 domains sampled). ICANN's September 14, 2010 *Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the Top 5 gTLDs* (**Privacy/Proxy Prevalence Study**) concluded that approximately 18 percent (± 2.0 percent) of domains in the top five gTLDs are likely registered using a privacy or proxy service.[19] Thus, the privacy/proxy rate found by the **WHOIS Registrant Identification Study** is statistically equivalent to the previous estimate. Data for the previous study was collected in the 2008-2009 time period, whereas the **WHOIS Registrant Identification Study** collected data in early 2012. Thus, there is no evidence to suggest that the usage of privacy and proxy services has changed over time.

We also broke this category into two components, Privacy Services and Proxy Services, based on the following definitions from the *Revised Terms of Reference for WHOIS Registrant Identification Studies*.[20]

- A **privacy service provider** offers alternate WHOIS contact information and mail forwarding services while not actually shielding the Registered Name Holder's identity.
- A **proxy service provider** registers a domain name on a third party's behalf and then licenses the domain name's use so that the provider's identity and contact information (and not the licensee's) is published in WHOIS.

Of the 320 domain registrants coded as Privacy/Proxy service providers, only 20 were determined to be privacy service providers. In other words, about 6 percent of domain registrants using a privacy or proxy service used a privacy service. ICANN's September 14, 2010 **Privacy/Proxy Prevalence Study** found a slightly larger percentage; approximately 9 percent of domain registrants using a proxy or privacy service used a privacy service. The difference between the percentages (3 percent) is not statistically significant at the 95 percent significance level.

Comparisons of the privacy and proxy services classifications between the two studies revealed apparent changes in the services offered by some providers. Four privacy/proxy service providers, which were determined to provide only proxy services in the current **WHOIS Registrant Identification Study**, were determined to provide both privacy and proxy services in the previous **Privacy/Proxy Prevalence**

---

[19] The September 28, 2009 draft *Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the Top 5 gTLDs* (http://www.icann.org/en/resources/compliance/reports/privacy-proxy-registration-services-study-28sep09-en.pdf) indicated that approximately 15 to 25 percent of domains in the top five gTLDs were likely registered using a privacy or proxy service. For the September 2010 final report (http://www.icann.org/en/compliance/reports/privacy-proxy-registration-services-study-14sep10-en.pdf), NORC refined the process for determining when WHOIS registrant information identified the use of a privacy or proxy service, and concluded that approximately 18 percent of sampled domains were registered using a privacy or proxy service.

[20] Op cit 6

Study.[21] There are 15 domains in the **WHOIS Registrant Identification Study** sample registered using one of these four service providers (4.7 percent of the 320 percent domain registered using a privacy or proxy service used a privacy service). While this may help to explain why the percentage of domains registered using a privacy service is slightly smaller in the current study, we do not have evidence to conclude that there is shrinkage of privacy service registration.

With such a small number of domains (20) in the privacy category, further analysis that attempts to cross-classify the privacy group with subject variables, such as potentially commercial activity, would not be meaningful. Therefore, our analyses combine privacy-registered and proxy-registered domains together.

Registrant's WHOIS Address Country/Region of the World

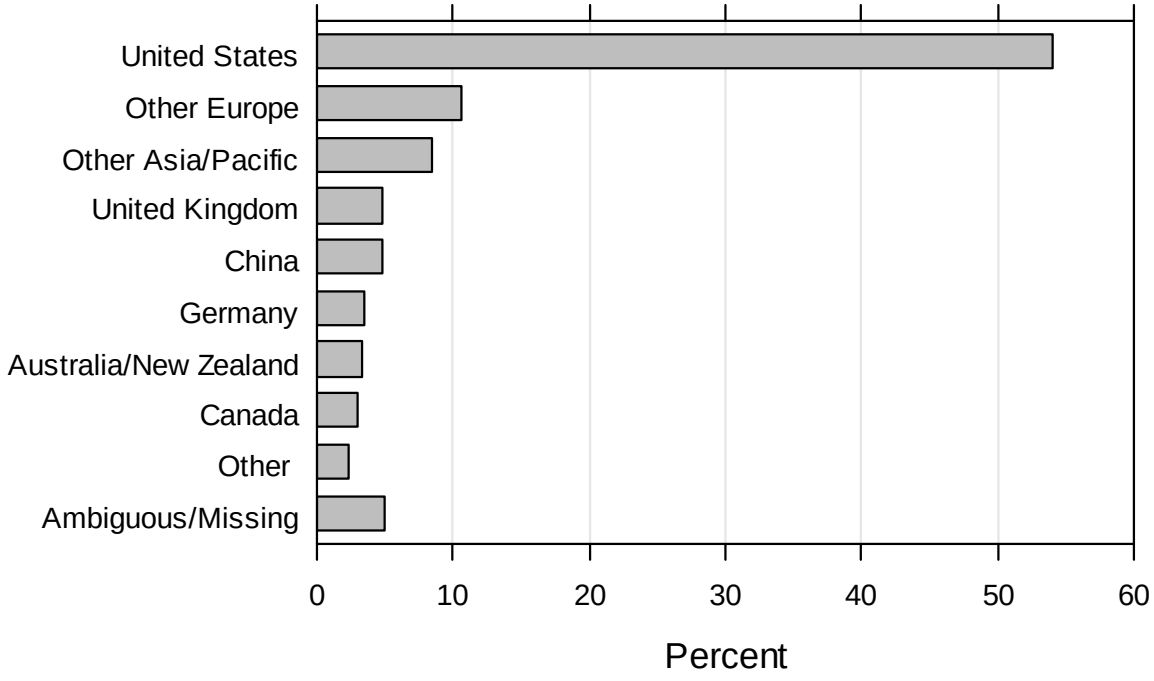
With respect to country information, we were able to identify the registrant's WHOIS address country for all but 82 of the domain names. For one domain name, there was conflicting information as to whether it was in Japan or Australia. For the remaining 81 missing registrant countries, the extracted WHOIS data was completely missing for 71 of the domains.[22] For the remaining 10 domains with missing registrant countries, the extracted data did not provide sufficient information to accurately code the registrant's country strictly based on the WHOIS data. Aside from these anomalies, 63 countries were represented in the sampled domains. For a complete list of the countries, and the number of domains associated with each country, see Appendix A, Section F.

Exhibit 3 shows the percentage of WHOIS registrant address country or region of the world. Countries with at least 50 domain names (United States, China, United Kingdom, Germany, and Canada) are shown in the chart. Other countries that appeared in the sample are grouped by region as follows: Other Europe, Other Asia/Pacific, Australia/New Zealand and Other (North America excluding the U.S. and Canada, South America, Caribbean Islands, and Africa).

[21] In all instances of **WHOIS Registrant Identification Study** sampled domain records, for which one of these four service providers was identified as the registrant, there is no licensee personally identifiable information in the registrant name or organization field. For the **Privacy/Proxy Prevalence Study**, some of the WHOIS records associated with these four services did have personally identifiable information in the registrant name or organization field, and were therefore classified as domains registered using a privacy service.

[22] At this point in time we can only speculate as to why these records were missing. It appears that in some cases, the domain expired between the time the domain name sample was selected and the time the NORC-BOT extracted the data. In other cases, a WHOIS record did not exist at the time of the NORC-BOT extraction, but it is possible a WHOIS record exists now. Because of the dynamic nature of the Internet environment it is difficult to determine why this happens.

Exhibit 3: Country/Region of the World from Registrant's WHOIS Address for Sampled Domains



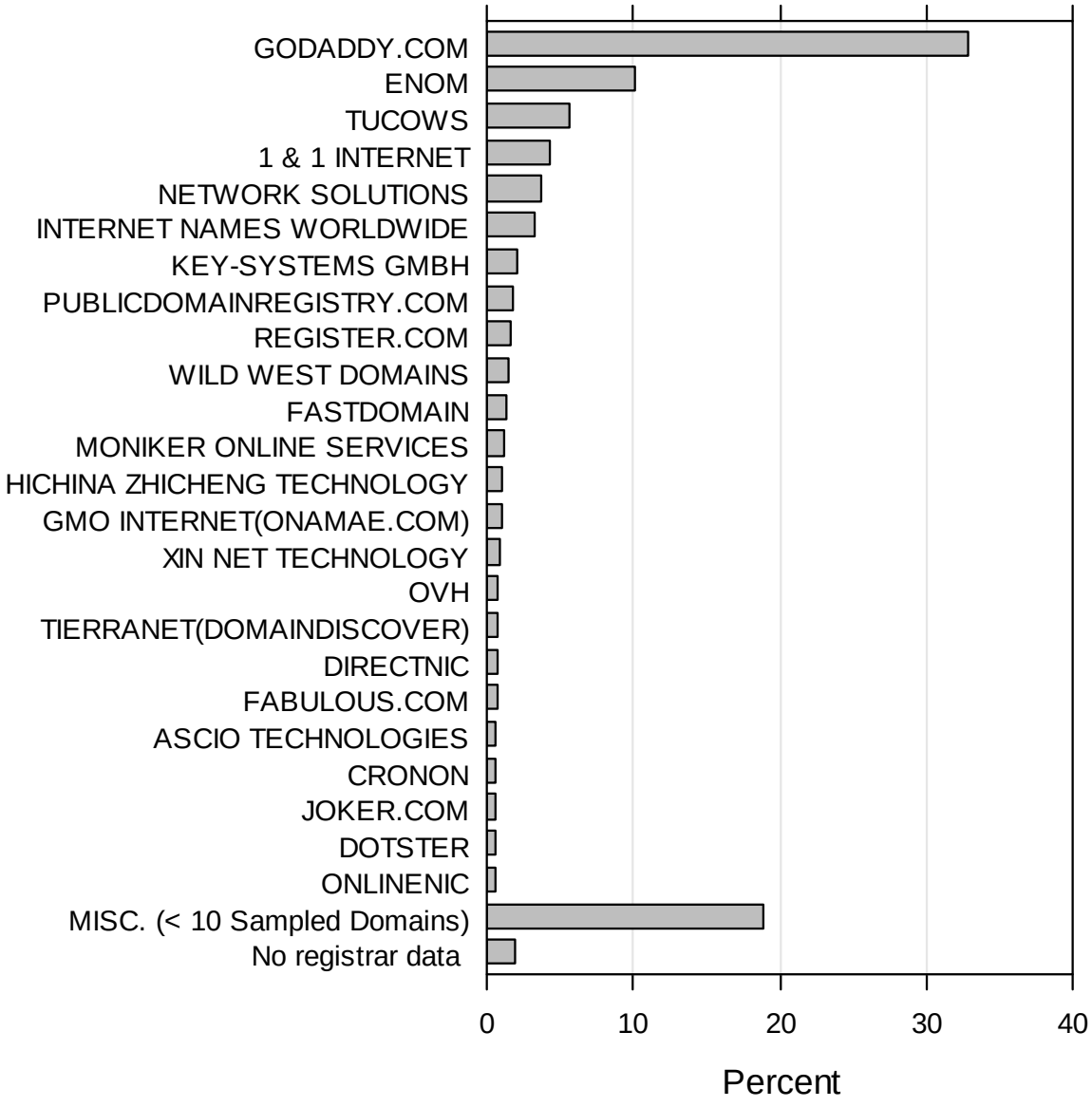
Other Europe = European countries other than the U.K. or Germany;
Other Asia/Pacific = Asian/Pacific countries other than China, Australia, or New Zealand
Other = countries in any of the following regions: North America excluding the U.S. and Canada, South America, Caribbean Islands, and Africa

The registrant country/region variable, without the Ambiguous/Missing category, is used in section 3 to further breakdown the main categories of interest: Apparent Registrant Type, Apparent Domain User Type and Potentially Commercial Activity.

Domain Registrars

Exhibit 4 shows the percentage of domains found by accredited registrars. Registrars with at least ten domains are shown individually in the chart. All remaining domains with registrar information are combined into one "miscellaneous" group. Registrar information could not be found for 32 sampled domains (2 percent). With such a large number of categories, further analysis based on registrar would not be statistically meaningful. Therefore, no additional cross-classifications are done between registrar and other variables of interest.

Exhibit 4: Domain Registrars Represented in the Domain Sample



2.3.2. Domain User Variables

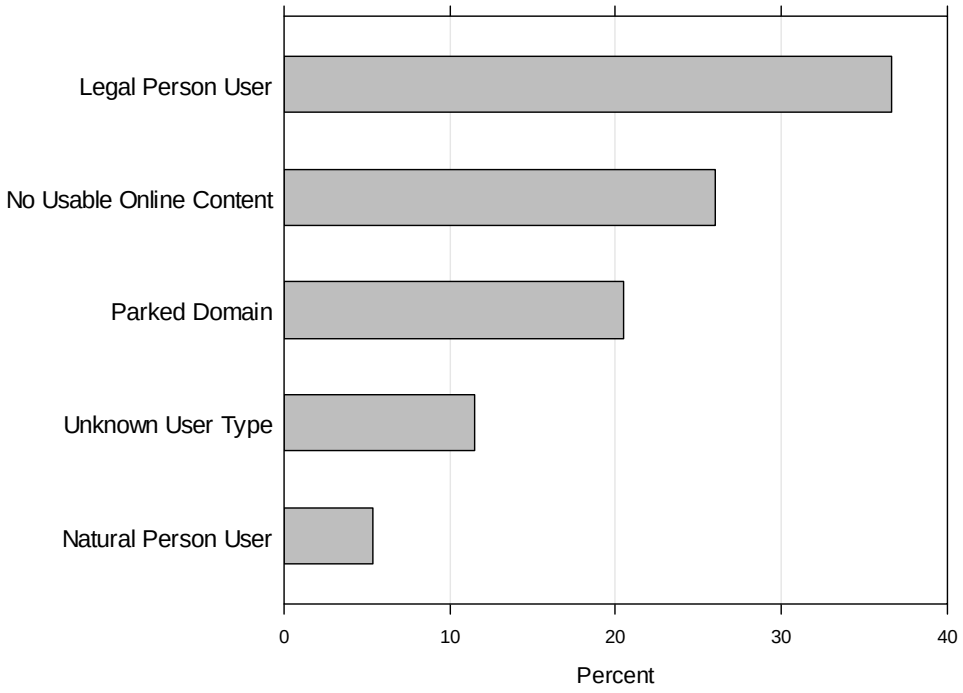
The Domain User is the beneficial user of a domain name. This is possibly different from the domain registrant. The entity that registered the domain may be the same as the entity using the domain, but there may also be no apparent relation. For example, norc.org is licensed and "used" by NORC at the University of Chicago. This can be confusing because everyone can be a "user" of norc.org, but for our purposes when we refer to the Domain User we are talking about NORC at the University of Chicago. The registrant of norc.org is NORC (the legal name for NORC at the University of Chicago), so in this case the registrant is the same as the Domain User, but this is not always the case.

Apparent Domain User Type

The Apparent Domain User Type was coded as to whether or not we could determine if the domain user could be considered a legal person or a natural person. As with Apparent Registrant Type, the domain user type was based on the evidence that we were able to discover; however, the web/FTP domain content downloaded by the NORC-BOT was the main information used to make this judgment. The domain user type could not be resolved for a number of domains. Sometimes this happened because a domain was "parked." In other situations, a domain website was offline or unreachable. The domain might also have basic HTML content that provides little-to-no usable content other than banner ads. Also a domain name might be used for non-web purposes. In other cases, domain web/FTP content that could be used for coding the domain user type was not available for the NORC-BOT to extract. There were still a number of domains that had web/FTP content, but it was not apparent whether the domain user was a legal or natural person. In such cases, the Apparent Domain User Type is "Unknown User Type."

Exhibit 5 provides a summary of Apparent Domain User Type for the sampled domains. Legal person users comprised 36.6 percent (± 2.4 percent) of all sampled domain user types. However, 5.4 percent (± 1.1 percent) are natural persons. Thus, there appear to be seven times as many legal person users than natural person users in our sample. However, for over half of our sample (56 percent), Apparent Domain User Type could not be determined based on retrieved web/FTP content. Of these indeterminate cases, many of the domains had no usable online content (26.0 ± 2.0 percent – which are referred to as "No Online Content" domains), or were parked (20.5 ± 2.0 percent). Only 11.5 percent (± 1.6 percent) of the domains had usable web/FTP content but that content was insufficient to determine Apparent Domain User Type.

Exhibit 5: Apparent Domain User Type for Sampled Domains



Further analyses of Apparent Domain User Type are summarized in Section 3.2. We review relationships between Apparent Domain User Type and each of the following variables: Apparent Registrant Type, Potentially Commercial Activity, Apparent Business Structure, Registrant's WHOIS Address Country/Region of the World, and the Domain User Relationship with the Registrant. Additional analysis of Apparent Domain User Type can be found in Appendix A.

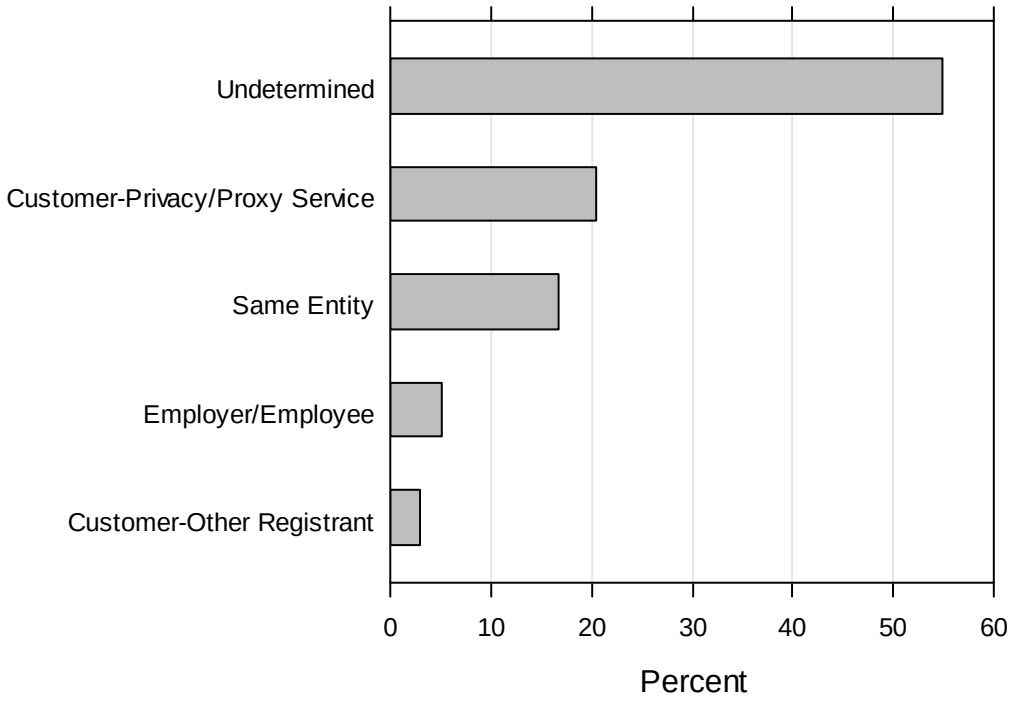
Domain User Relationship to Registrant

In additional to the domain user type, we also investigate the relationship between the domain user and the registrant. We settled on five broad categories for the relationship:

1. Domain user is the same as the registrant,
2. Domain user is a customer of a Privacy/Proxy service,
3. Domain user is a customer of the registrant (e.g. a web developer, development, or consulting company registered the domain, but not a Privacy/Proxy service—referred to as Other Registrant),
4. Domain user is an employer/employee of the registrant, or
5. Unknown relationship—there was not enough evidence present in the domain content to classify the relationship into one of the four other categories.

Exhibit 6 provides a summary of domain user's relationship to the registrant for the sampled domains. Approximately 55 percent (± 2.4 percent) of the relationships in the sampled domains are Undetermined. In section 3.2 we investigate the relationship between Apparent Domain User Type and Domain User Relationship to the Registrant. The percentage of domains with an undetermined relationship is higher than the overall percentage when the Apparent Domain User Type could not be determined (i.e. No Online Content, Parked Domain, or Unknown User Type—see Exhibit 19). However, there are cases where the relationship is undetermined for apparent legal person and apparent natural person user types.

Exhibit 6: Domain User Relationship to Registrant for Sampled Domains



Domain User Relationship to Registrant is used in Section 3 to further breakdown the main categories of interest: Apparent Registrant Type, Apparent Domain User Type and Potentially Commercial Activity.

Apparent Business Structure

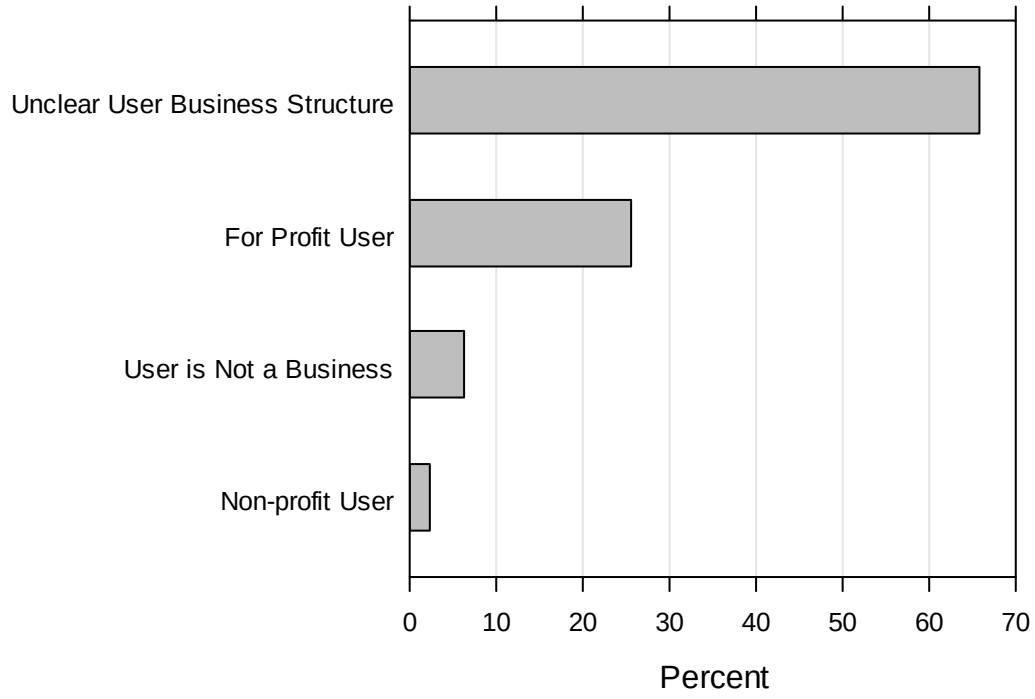
We also developed classes for the business structure and function of the domain users that appeared to have some sort of potentially commercial activity represented. Basing our initial classes on standard business classifications (partially based on classifications at digitalenterprise.org/models/models.html), we attempted to distinguish between corporations and smaller entities such as partnerships and sole proprietorships in the business structure variable. The same process in the business function variable yielded fourteen codes; among these were traditional functions such as Enterprise, Retail, Non-Profit, Consultancy, as well as newer digital functions, such as Utility and Domain Parking. Initial coding

attempts revealed the difficulty of making clear determinations within these variables; there were 940 domains (58.9 ± 2.4 percent) whose business structure and function could not be discerned by our coding team. This includes all of the 416 sampled domains that had no online content, the 328 parked sample domains, and 182 of the 183 sampled domains with unknown domain user type. We found that while some designations, such as Corporation (in structure) and Enterprise (in function) were often prominently stated or easily inferred from web content, many other determinations were either impossible to make or admitted ambiguity with overlapping possible designations. For example, while it was often clear that a website promoted a small business, it would not be clear whether this business was a partnership or a sole proprietorship. In the business function variable, many businesses often fit within multiple categories and required rather fine distinctions to be made.

We addressed these difficulties by completing two rounds of coding for each sampled domain, with extensive training sessions before each to explicate the common distinctions we expected our coders to make. In both rounds, coders were given a standard code frame to reference in their coding, while also having the latitude to note unique circumstances for each domain. Coders analyzed not only the downloaded web content for references to business structure and indications of function, but also employed third party services such as Accurint and LinkedIn to provide supplemental information or corroboration for codes. Online translators (including Google's translation function) were used to decipher foreign language pages. After this process, analysts performed adjudication of these two sets of codes in order to reconcile discrepancies into a set of codes with more uniformly applied rules, while also taking into account the special circumstances noted by coders. Adjudicators also developed a set of generic structure and function codes to consolidate the myriad designations into more abstract categories (thus, codes such Partnership and Sole Proprietorship in the Business Structure variable were consolidated into "Small Business" code). Throughout the process, we emphasized that indecipherable cases should be coded as one of the various "Unknown" codes to maintain the high-quality nature of the data. Keyword searches were used to help with the coding, but there were not enough keywords identified that made the coding process fully automated. Some manual review was needed to check and complete business structure coding. The use of keywords in the coding process might prove effective in future studies. More research is needed to make a conclusion one way or another.

We settled on four general categories to denote business structure for analysis purposes. These Apparent Business Structure codes are summarized for the sampled domains in **Exhibit 7**. Approximately 66 percent (± 2.3 percent) of all sampled domains had an unclear business structure, while about 26 percent (± 2.1 percent) were assessed as for profit businesses. Appendix A, Section D provides more background on the Business Structure variable, as well as tabular analyses summaries.

Exhibit 7: Domain User – Apparent Business Structure for Sampled Domains



Apparent Business Structure is used in Section 3 to further breakdown the main categories of interest: Apparent Registrant Type, Apparent Domain User Type and Potentially Commercial Activity.

2.3.3. Domain Content Variables

Domain content refers the information that was downloaded from each sampled domain's primary HTTP/FTP server, and the types of apparent activities taking place at that site. In particular, we focus on the types of potentially commercial activities suggested by the domain content, and whether or not there is an appearance of allegedly illegal or harmful activities associated with the domain.[23] For the later, we incorporate response codes from DNS BlackLists (DNSBL) to help make the determination of allegedly illegal or harmful activities.

Potentially Commercial Activity

One of the key domain content variables is Potentially Commercial Activity. Per the GNSO Council's request, we attempted to categorize all observed monetary activities that in some countries might be legally considered "commercial activities," documenting a broad range of potentially commercial activities to enable multiple post-study interpretations that apply varied legal definitions. We looked for

[23] NORC understands that ICANN has commissioned a study to explore privacy/proxy abuse. The "WHOIS Privacy and Proxy Abuse Study" is exclusively focused on finding domains engaged in allegedly illegal or harmful activity. The Registrant ID Study does not have this focus; however an attempt was made to categorize any allegedly illegal or harmful activity that was apparent in the domain sample.

evidence of e-commerce, collection of membership dues for online content or offline content, promotional material content, banner ads, and pay-per-click ads. If any of these monetary activities were perceived to take place at a domain, we considered this domain to have Potentially Commercial Activity.

Exhibit 8 is a summary of Potentially Commercial Activity for sampled domains. Note that a sampled domain could show evidence of one or more of the activities of interest. Thus, the counts of detected activities do not add across activity categories.

Exhibit 8: Potentially Commercial Activity Observed for Sampled Domains

Commercial Activity Variable	Detected	Percent	Margin of Error (±)
Promotional Content	511	31.9	2.3
Promotional Content (Offline)	295	18.4	1.9
Promotional Content on Host	139	8.7	1.4
Promotional Content (Online)	93	5.8	1.1
Pay-Per-Click Ads	483	30.2	2.2
Pay-Per-Click Ads (Non-Host)	469	29.3	2.2
Host Pay-Per-Click Ads	61	3.8	0.9
Banner Ads	306	19.1	1.9
Host Banner Ads	202	12.6	1.6
Third Party Banner Ads	104	6.5	1.2
E-Commerce	111	6.9	1.2
Membership Dues	83	5.2	1.1
Membership (Offline Content)	56	3.5	0.9
Membership (Online Content)	28	1.8	0.6
Any Potentially Commercial Activity	**905**	**56.6**	**2.4**

Exhibit 8 provides information to answer the GAC question: ***What is the percentage of domain name uses that are commercial versus non-commercial?***

Approximately, 57 percent (± 2.4 percent) of all sampled domains were perceived to have potentially commercial activity.

Because pay-per-click ads were found so frequently that they dominated this variable, we completed our analysis with and without pay-per-click ads to enable both interpretations of potentially commercial activity. If pay-per-click ads were not considered potentially commercial activity, then the number of sampled domains with potentially commercial activity would

decrease to 717. In this case, approximately, 45 percent (± 2.4 percent) of all sampled domains have potentially commercial activity.

We also considered measuring the degree of potentially commercial activity. Attempts were made to count the number of potentially commercial activity data elements found at a domain, e.g. count the number of different pay-per-click ads embedded in web pages. This proved difficult to do regardless of whether a manual or automated process was used. The quality and consistency of such counts were not reliable. Therefore we decided to only consider whether or not a domain was perceived to have any kind of potentially commercial activity.

Additional analyses of Potentially Commercial Activity (including pay-per-click ads) are described in Section 3.3. Furthermore, the variable is used in other Section 3 subsections to further breakdown the main categories of interest: Apparent Registrant Type, Privacy/Proxy Use, and Apparent Domain User Type. Appendix A, section C includes additional results when pay-per-click ads are not considered Potentially Commercial Activity.

Allegedly Illegal or Harmful Activities and Explicit Sexual Content

We also looked for allegedly illegal or harmful activities using manual and automated processes. The manual process called for coders to make a judgment based on observed HTTP/FTP domain content. Additionally, the presence of explicit sexual imagery was recorded during the manual process. Only 18 domains were classified as having allegedly illegal or harmful activities (1.1 ± 0.5 percent) and 16 domains contained explicit sexual content (1.0 ± 0.5 percent). Further cross-classified analysis of these data for the purpose of determining if these two behaviors are more likely among certain subgroups (of key study variables such as Privacy/Proxy Services) is questionable given the small number of observations that apparently exhibit the behavior. Because there is interest in the ICANN community in such a drill-down, we still carried out analyses to see if these two behaviors were more likely among certain subgroups. Appendix A, Section H contains detailed analysis. As expected, there are no statistically significant results.

Exhibit 9 shows the results of comparing the presence of Potentially Commercial Activity within two behavior groups: Allegedly Illegal or Harmful Activities and Explicit Sexual Images. For both behavior groups, a higher percentage of domains are classified as Potentially Commercial Activity. However, the differences between domains with and without Potential Commercial Activity are insignificant given the small number of observations in each behavior group (16 and 18).

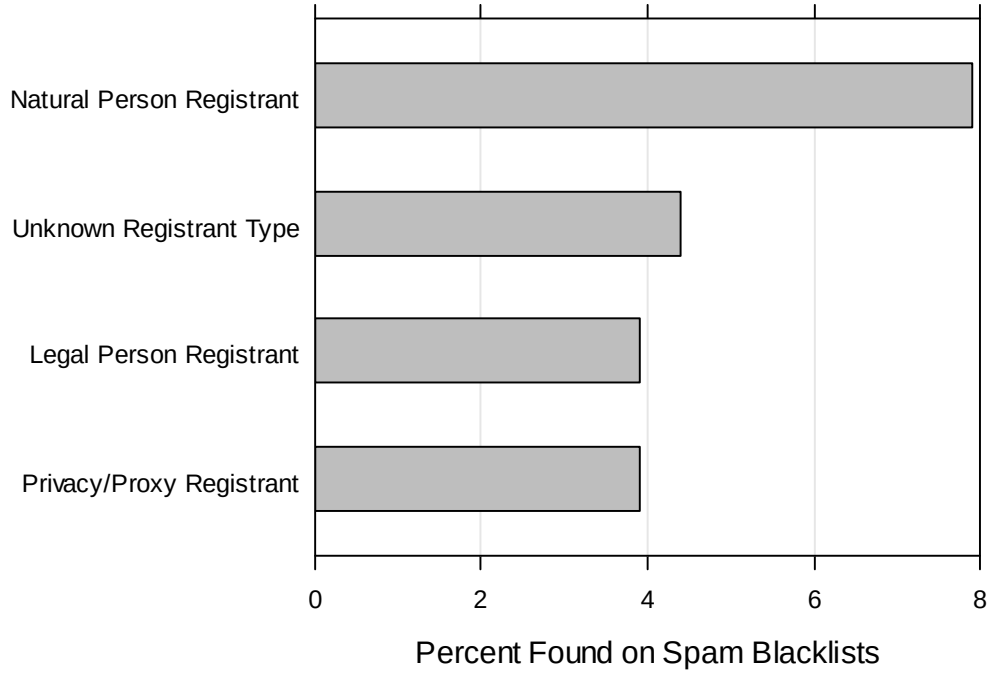
Exhibit 9: Potentially Commercial Activity Within Behavior Groups

Behavior Variable	Percent Exhibiting Behavior: No Potentially Commercial Activity	Percent Exhibiting Behavior: Potentially Commercial Activity
Allegedly Illegal or Harmful Activity	0.8	1.5
Explicit Sexual Images	0.9	1.2

The automated process for detecting allegedly illegal or harmful activities used scans of blacklists (DNSBL lists). NORC created a scoring system for blacklists, and sample domains found on top rated lists where identified. Overall, 141 sampled domains were found on at least one blacklist (8.8 ± 1.4 percent). Whitelists, which are lists used to exempt a domain or URL from black-listing, were also scanned. We found 204 sampled domains on whitelists (12.8 ± 1.6 percent); 13 of which were also found on a blacklist. Appendix A, Section I contains the blacklist analysis results, and Section J contains the whitelist analysis results.

The results of these analyses are mixed. A breakdown of blacklisting by whether or not potentially commercial activity is present does not produce statistically significant results. On the other hand, breakdowns by Apparent Registrant Type and Apparent Domain User Type does; especially for spam monitoring blacklists. **Exhibit 10** shows the percent of domains found on spam blacklists for the four types of Apparent Registrant Type. Domains of natural person registrants are almost twice more likely to appear on spam blacklist than the other Apparent Registrant Types; albeit the percentage is just 8 percent.

Exhibit 10: Presence on Spam Blacklists Within Apparent Registrant Type



The interested reader should consult Appendix A to review the rest of the blacklist and whitelist analyses. The remainder of this summary report will concentrate on drill-downs into Apparent Registrant Type (including a focus on Privacy/Proxy Services), Apparent Domain User Type, and Potentially Commercial Activity.

3: Relationship Analysis Results

In this section, we drill-down into Apparent Registrant Type, Apparent Domain User Type, and Potentially Commercial Activity variables by cross-classifying each variable with the other two, as well as cross-classifying each with Registrant's WHOIS Address Country/Region of the World, Domain User Relationship to Registrant, and Apparent Business Structure. This provides a means to observe the occurrence of cross-classifying variables categories relative to each main variable category, and it provides a comparison of each cross-classifying category across the main variable categories. **Appendix A: Exploratory Analysis Report** provides additional details related to the results of statistical tests for association between cross-classified variables.

As previously noted, this is an exploratory study that was designed to provide information to answer the GAC questions in a statistically accurate way.[24] In this sense, a sample of 1,600 domains was selected from the top five gTLDs assuming that most questions of interest would pertain to domain subgroups with (approximately) 400 or more subgroup members. For example the subgroups legal person user and domains with commercial use each have over 400 domains (586 and 905 subgroup member domains, respectively). Thus, the GAC questions related to the relative proportion of Privacy/Proxy service use can be answered with estimates that meet the statistical accuracy goal.

Another aspect of an exploratory study is observing whether or not there is an association (dependence) between variables of interest. For example, we look to see what associations may exist between Apparent Registrant Type and Apparent Domain User Type. NORC's main tool for statistically establishing associations between variables is the Chi-squared Test of Association.[25] The null hypothesis of this statistical test is that the two categorical variables are independent (not associated). If the observed chi-squared test statistic, which is based on the difference between observed and expected cross-classified frequencies, is unusually large assuming the null hypothesis of independence is true, then we infer that independence assumption is suspect and conclude that the two categorical variables are associated (dependent upon one another). Appendix A describes the details and results of the tests of association we performed. In this section, we review some of the variable relationships where statistically significant associations are found, and we use visual exploration to better understand the associations.

[24] For this study, a proportional estimate is considered statistically accurate if its margin of error at the 95 percent confidence level is no more than plus/minus 5 percentage points.

[25] Greenwood, P.E., Nikulin, M.S. (1996) A guide to chi-squared testing. Wiley, New York. ISBN 0-471-55779-X

In what follows, graphics are used to illustrate cross-classification percentages. The term main variable refers to Apparent Registrant Type, Apparent Domain User Type, or Potentially Commercial Activity. The cross-classification variables are the three main variables, as well as, Registrant's WHOIS Address Country/Region of the World, Domain User Relationship to Registrant, and Apparent Business Structure.

Most of the graphics are organized like **Exhibit 11**. Each main variable category is represented by a separate panel in the graphic, and a row in each panel corresponds to the relative percentage of sampled domains for a category of the cross-classification variable. The cross-classification variable's categories are sorted, from highest to lowest percentage, based on the overall sample percentages, which is mark with "+" in the graph. Associations between the two variables in the graphics can be explored in two ways: 1) within each panel (main variable category), and 2) across each panel.

Because the ordering of the cross-classification variables categories is done using the overall sample results, differences between the overall results and those within a main variable categories are usually apparent. Thus, when the overall cross-classification variable ranking changes within a main variable category, association between the variables becomes apparent. By comparing cross-classification category relative percentages across panels, we can also judge the effect the main variable has on the cross-classification variable.

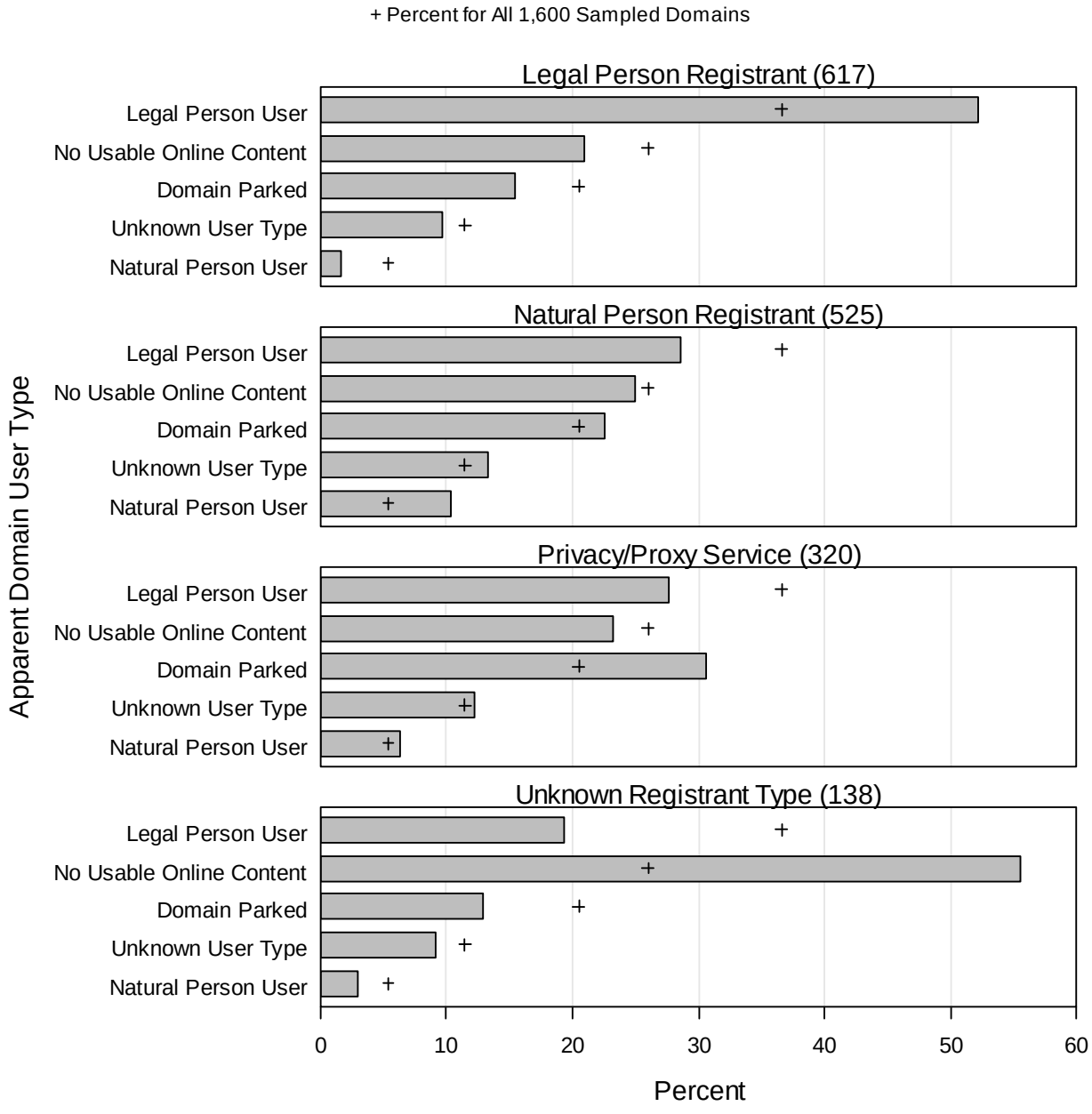
For each of the analysis summaries presented in this report, we point out the apparent reasons for the statistically detected associations between the cross-classified variable and the main variable. Additionally, we provide relative percentage estimates along with the margins of error at the 95 percent confidence level.

Each reader of this report may notice differences between variables of interest that are not discussed in this summary report. To help readers determine if the differences are meaningful, sample size information is provided in the graphic labels for the main variable categories. For example, in the label "Legal Person Registrant (617)," the number of sampled domains in the legal person registrant category of Apparent Registrant Type is 617. The sample size information, in part, helps to explain the size of the margin of error for some of the estimates. If two samples of different sizes provide the same proportional estimate value, then the margin of error will be smaller for the larger of the two samples. Thus, estimates based on large samples, generally provide more accurate estimates. Additional analyses are also provided in Appendix A.

3.1. Apparent Registrant Type

Apparent Domain User Type

Exhibit 11: Apparent Domain User Types Within Apparent Registrant Type Categories



Comparing the relative percentage of Apparent Domain User Types in **Exhibit 11** both within and across Apparent Registrant Type, we can observe the following associations between these two variables.

- As might be expected, domain names registered by legal persons were more likely to be used by legal persons—52.2 ± 3.9 percent, as compared to the entire sample's 36.6 percent. Domain User Type ranking for domains registered by legal persons is consistent with ranking across the entire sample—that is, as shown in **Exhibit 5**, legal person (36.6 percent), no online content (26.0 percent), domain parked (20.5 percent), unknown user type (11.5 percent), and natural person (5.4 percent).

- Similarly, we found that domain names registered by natural persons were more likely to be used by natural persons—10.4 ± 2.6 percent, as compared to the entire sample's 5.4 percent. Here again, Domain User Type ranking for domains registered by natural persons is consistent with overall sample ranking.

The overall sample ranking does not hold within the other Apparent Registrant Types.

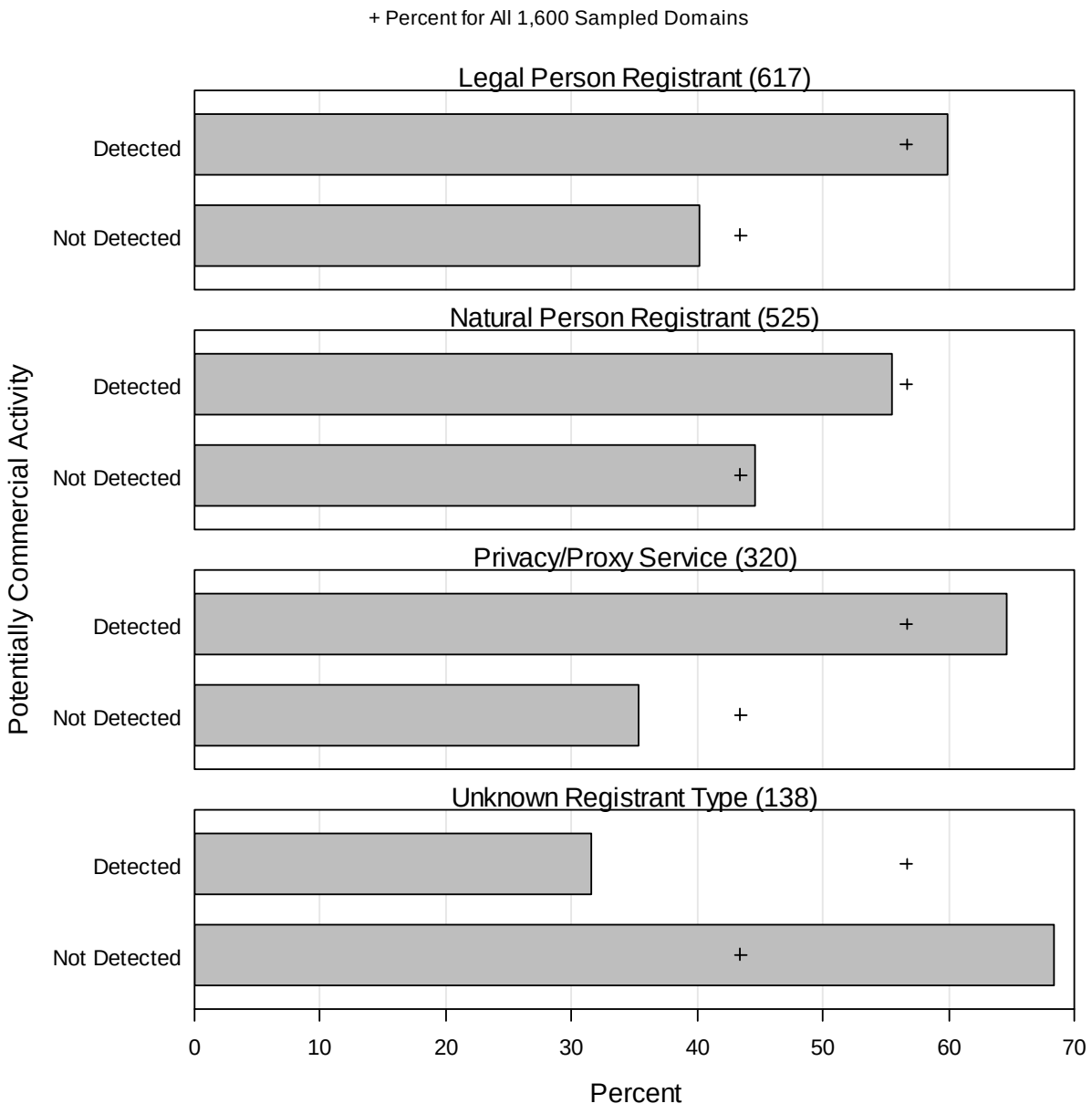
- Domain names registered by entities that could not be classified using WHOIS were more likely to not have usable online web/FTP content—55.5 ± 8.2 percent, as compared to the overall sample's 26.0 percent. This suggests that someone who encounters difficulty using WHOIS data to identify a registrant may also be likely to have trouble identifying that domain's user based on web/FTP content.

- Domain names registered using a Privacy/Proxy service were more likely to be parked—30.7 ± 5.0 percent, as compared to the overall sample's 20.5 percent. Coupled with domains for which there was no usable online content, over half of domain names registered by Privacy/Proxy services appear to be domain names that possibly are held for resale or other uses that do not involve web/FTP content.

Additional exploration of the relationship between Apparent Registrant Type and Apparent Domain Type is provided in section 3.2.

Potentially Commercial Activity

Exhibit 12: Detection of Potentially Commercial Activity Within Apparent Registrant Type Categories



Comparing the relative percentage of Potentially Commercial Activity in **Exhibit 12** both within and across Apparent Registrant Type, we can observe the following associations between these two variables.

- As might be expected, domain names registered by legal persons were slightly more likely to be used for some kind of potentially commercial activity—59.9 ± 3.9 percent. Statistically speaking, the detection percentage is the similar to the entire sample's 56.6 percent.

- Similarly, domain names registered by natural persons were equally as likely to be used for some kind of potentially commercial activity as the overall sample—55.4 ± 4.3 percent, as compared to the entire sample's 56.6 percent.

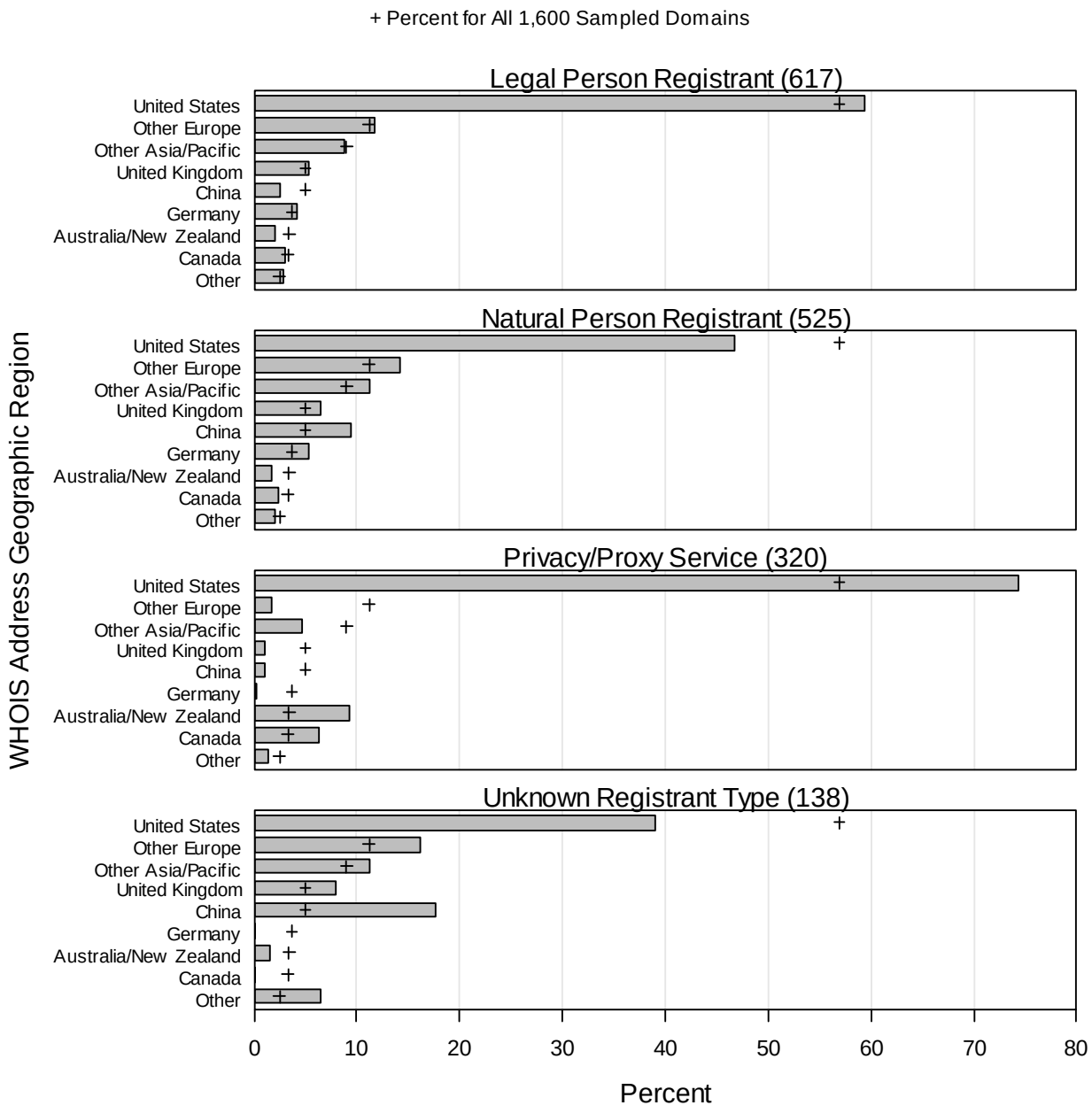
- Domain names registered by a privacy/proxy service were the most likely to be used for some kind of potentially commercial activity—64.6 ± 5.2 percent.

- Potentially commercial activity was less likely to be found for domain names with unknown registrant type—31.6 ± 7.7 percent, as compared to the overall sample's 56.7 percent. This suggests that registrants of domains that are not concerned with providing usable WHOIS data tend to not use domains for potentially commercial activity.

Additional exploration of the relationship between Apparent Registrant Type and Potentially Commercial Activity is provided in section 3.3.

Registrant's WHOIS Address Country/Region of the World

Exhibit 13: Country/Region of the World from Registrant's WHOIS Address Within Apparent Registrant Type Categories



Other Europe = European countries other than the U.K. or Germany;
Other Asia/Pacific = Asian/Pacific countries other than China, Australia, or New Zealand
Other = countries in any of the following regions: North America excluding the U.S. and Canada, South America, Caribbean Islands, and Africa

Comparing the relative percentage of Registrant's WHOIS Address Country/Region of the World in **Exhibit 13** both within and across Apparent Registrant Type, we can observe the following associations between these two variables.

- Legal persons that register domain names are slightly more likely to have WHOIS addresses in the U.S.—59.4 ± 3.9 percent. This is statistically equivalent to the overall sample percentage of 56.9 percent.

    - The ranking of legal person Registrant's WHOIS address county/region of the world is almost consistent[26] with overall sample ranking—that is U.S. (56.9 percent), Other Europe (11.2 percent), Other Asia/Pacific (9.0 percent), United Kingdom (5.0 percent; tied with China), China (5.0 percent), Germany (3.7 percent), Australia/New Zealand (3.4 percent), Canada (3.3 percent), Other(2.5 percent).[27]

    - An exception here is that legal persons that register domain names are statistically less likely to have a China address—2.6 ± 1.2 percent compared with the 5.0 percent for the entire sample.

- Natural persons that register domain names are less likely to have addresses in the U.S.—46.0 ± 4.3 percent as compared to the entire sample's 56.9 percent. Again, natural person registrant's WHOIS address country/region rankings are consistent with the overall sample with China being an exception of note—9.3 ± 2.5 percent.

The overall sample ranking does not hold within the other Apparent Registrant Types.

- Domain names registered by privacy/proxy services are most likely to be registered with a WHOIS address in the U.S.—74.3 ± 4.8 percent. Australia/New Zealand—9.3 ± 3.2 percent—and Canada—6.3 ± 2.7 percent—also have higher than expected percentages compared to the entire sample. In general, the other countries/regions have about the same or lower percentages of privacy/proxy registration as compared to the entire sample.

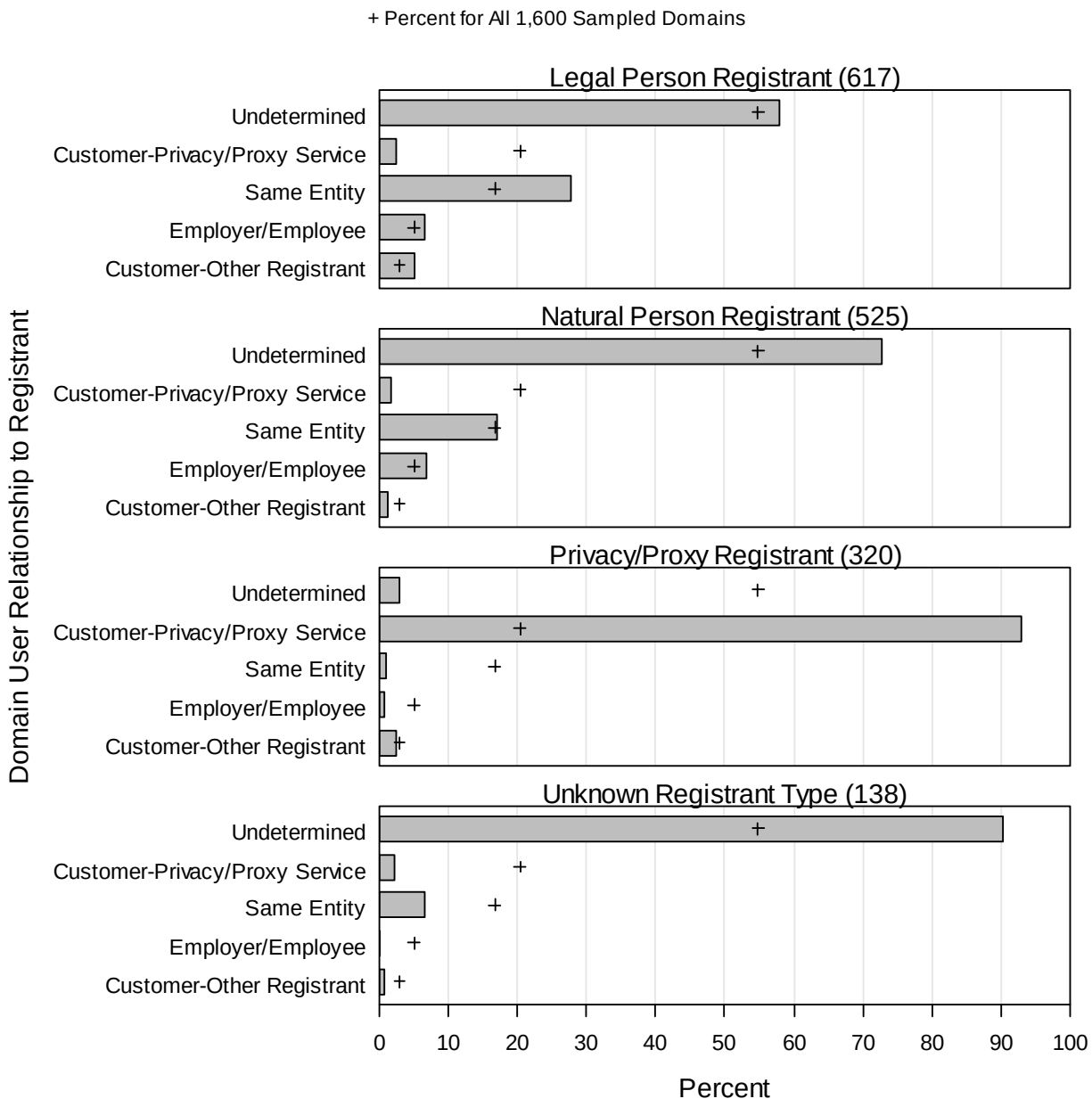
- Domain names registered by entities that could not be classified using WHOIS were more likely to have a WHOIS registrant address outside of the U.S. In particular, the percentage registered with a Chinese address is higher than expected—17.7 ± 9.3 percent, as compared to the entire sample's 5.0 percent. It is possible that our unfamiliarity with naming conventions in China made it difficult to determine the registrant type.

[26] There are some slight differences between the legal person registrant rankings versus the overall sample; however most of the differences are not statistically significant.

[27] Note that the overall sample percentages are similar to those shown in **Exhibit 3**. However, the percentages are relative to the 1,518 WHOIS records that did not have ambiguous or missing country address information.

Domain User Relationship to Registrant

Exhibit 14: Domain User Relationship to Registrant Within Apparent Registrant Type Categories



Comparing the relative percentage of Domain User Relationship to Registrant in **Exhibit 14** both within and across Apparent Registrant Type, we can observe that the overall sample ranking does not hold within any of the Apparent Registrant Types—that is, as shown in **Exhibit 6**, Undetermined (54.8 percent), Customer-Privacy/Proxy Service (20.4 percent), Same Entity (16.8 percent), Employer/Employee (5.1 percent), and Customer-Other Registrant (3.0 percent).

- Domain names registered by legal persons were more likely to also be used by that legal person—27.8 ± 3.5 percent, as compared to the entire sample's 16.8 percent. However, we were not able to determine the user/registrant relationship for a majority of legal person registrants—57.9 percent ± 3.9 percent, which is similar to the entire sample's undetermined relationship rate of 54.8 percent.

- Domain names registered by natural persons were more likely to have undetermined domain user/registrant relationships—72.5 ± 3.9 percent, as compared to the entire sample. Domain names registered by natural persons were also less likely to be used by customers of a privacy/proxy service—1.9 ± 1.2 percent, as compared to the entire sample's 20.4 percent. This makes sense; if a proxy service is used to register the domain then the registrant will not be a natural person. For the other user/registrant relationship types, the relative percentages are similar to the overall sample percentages.

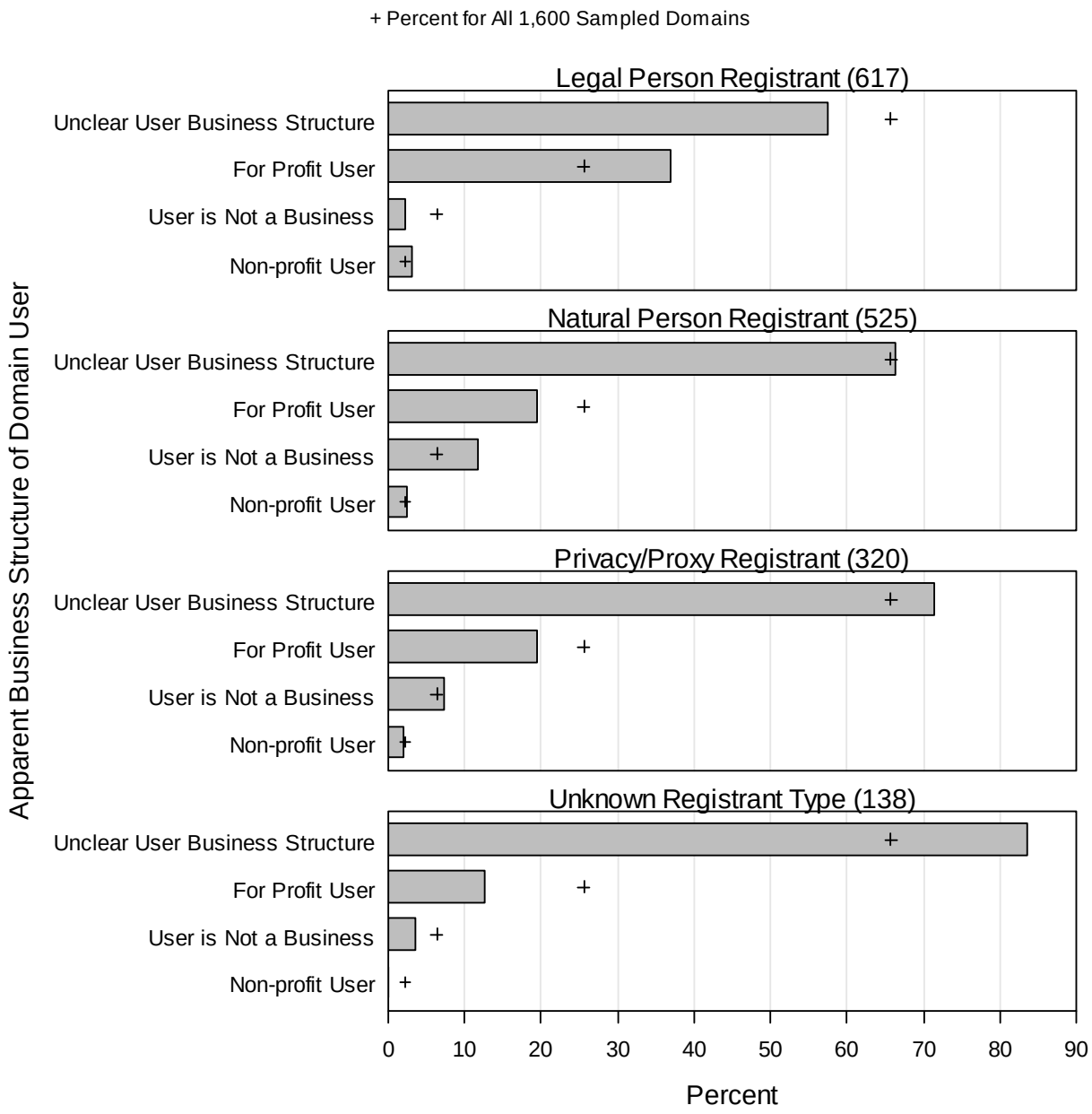
- Not surprisingly, when a domain name is registered through a privacy/proxy service, it is almost always the case the user/registrant relationship is customer of a privacy/proxy service—92.8 ± 2.8 percent.

    - This relative percentage is not 100 percent because NORC's coding of this variable used the identity of the entity that presumably contracted a privacy service to register the domains. In such cases, the registered name holder's identity was not shielded, and we could determine the relationship with the domain user.

- As might be expected, domain names registered by entities that could not be classified using WHOIS were the most likely to have undermined user/registrant relationships—90.3 ± 4.9 percent, as compared to the entire sample's 54.8 percent.

Apparent Business Structure

Exhibit 15: Apparent Business Structure of Domain User Within Apparent Registrant Type Categories



Comparing the relative percentage of the Domain User's Apparent Business Structure in **Exhibit 15** both within and across Apparent Registrant Type, we can observe the following associations between these two variables.

- Domain names registered by legal persons were more likely to be used by a for-profit entity—39.9 ± 3.8 percent, as compared to the entire sample's 25.6 percent. Business Structure ranking

for domains registered by legal persons is consistent with ranking across the entire sample—that is, as shown in **Exhibit 7**, unclear business structure (65.7 percent), for-profit (25.6 percent), not a business (6.4 percent), and non-profit (2.3 percent).

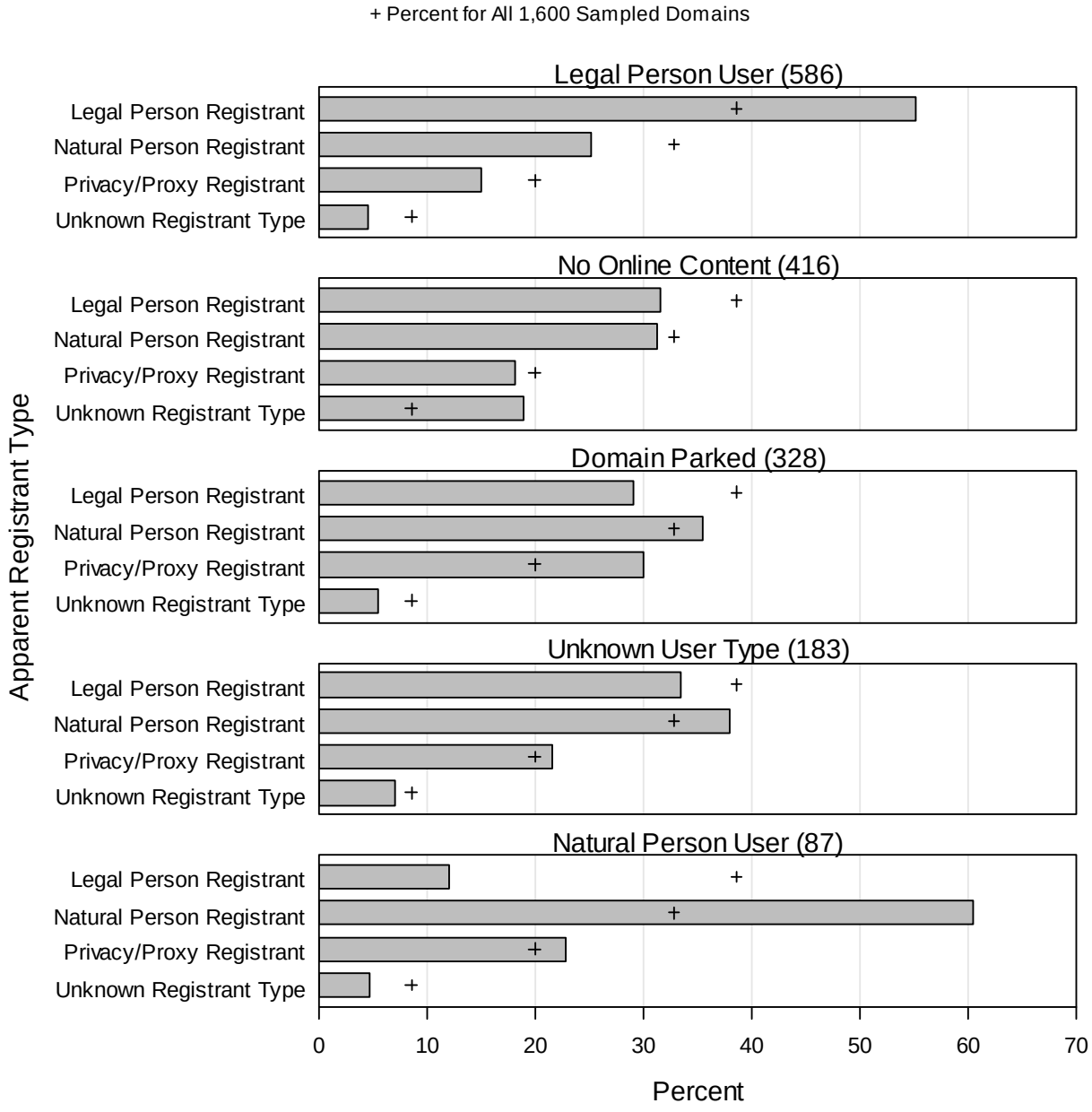
- As might be expected, domain names registered by natural persons were more likely to be used by a non-business entity—11.8 ± 2.8 percent, as compared to the entire sample's 6.4 percent. Here again, Business Structure ranking for domains registered by natural persons is consistent with overall sample ranking.

- The domain user's business structure was predominately unclear for all Apparent Registrant Types, and (perhaps not surprisingly) it was most prevalent for domains that could not be classified using WHOIS information—83.6 ± 6.1 percent, as compared with the entire sample's 65.7 percent. This suggests that someone who encounters difficulty using WHOIS data to identify a registrant may also be likely to have trouble using web/FTP content to determine whether the domain user is a for-profit business, a non-profit entity, or not a business at all.

3.2. Apparent Domain User Type

Apparent Registrant Type

Exhibit 16: Apparent Registrant Type Within Apparent Domain User Type Categories



Comparing the relative percentage of Apparent Registrant Types in **Exhibit 16** both within and across Apparent Domain User Type, we can observe the following associations between these two variables.

- As might be expected, domain names used by legal persons were more likely to be registered by legal persons—55.1 ± 4.0 percent, as compared to the entire sample's 38.6 percent. Apparent

Registrant Type ranking for domains used by legal persons is consistent with ranking across the entire sample—that is, as shown in **Exhibit 5**, legal person (38.6 percent), natural person (32.8 percent), privacy/proxy service (20.0 percent), and unknown registrant type (8.6 percent).

The overall sample ranking does not hold within the other Apparent Domain User Types.

- Not surprisingly, domain names used by natural persons were more likely to be registered by natural persons—60.4 ± 10.2 percent, as compared to the entire sample's 32.8 percent. Whereas, domain names used by natural persons were less likely to be registered by legal persons—10.1 ± 6.8 percent, as compared to the entire sample's 38.6 percent.

- Domain names without usable online web/FTP content were more likely to be registered by entities that could not be classified using WHOIS—18.9 ± 3.8 percent, as compared to the overall sample's 8.6 percent. This suggests that someone who encounters a domain without web/FTP content may have more difficulty using WHOIS data to identify a registrant than for other domain user types.

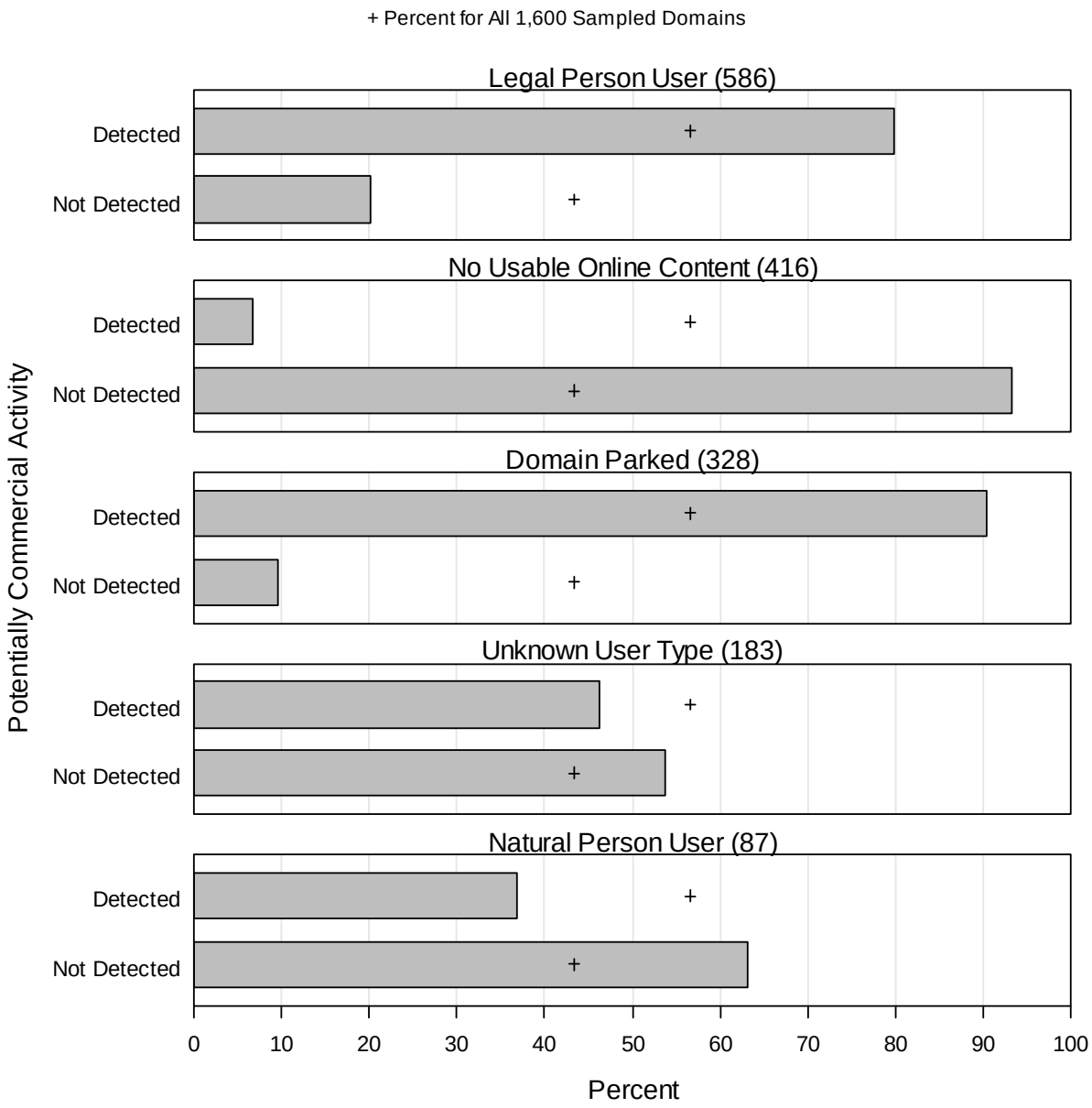
- Domain names that possibly are held for resale or other uses that do not involve web/FTP content (parked domains) were more likely to be registered by privacy/proxy services—30.0 ± 5.0 percent, as compared to the overall sample's 20.0 percent.

Exhibit 16 also provides information to answer the GAC question: ***What is the relative percentage of Privacy/Proxy use among legal persons?***

In our random sample of 1,600 domains, 586 are used by apparently legal persons. Of these domains, 15.1 percent (± 2.9 percent) were registered using a privacy or proxy service.

Potentially Commercial Activity

Exhibit 17: Detection of Potentially Commercial Activity Within Apparent Domain User Categories



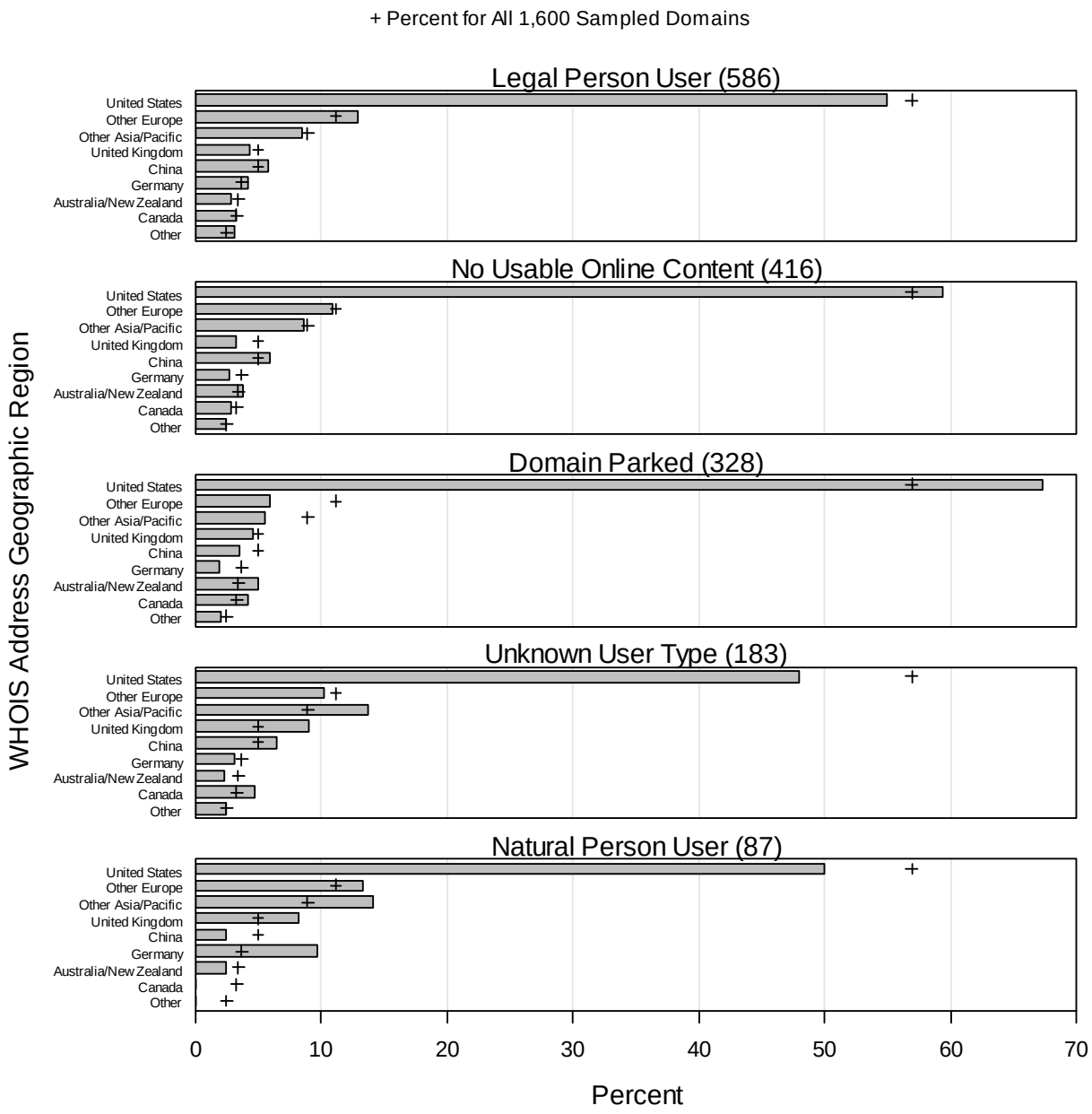
Comparing the relative percentage of Potentially Commercial Activity in **Exhibit 17** both within and across Apparent Domain User Type, we can observe the following associations between these two variables.

- As might be expected, domain used registered by legal persons were more likely to be used for some kind of potentially commercial activity—79.8 ± 3.2 percent, as compared to the entire sample's 56.6 percent.

- As possibly anticipated, domains used by natural persons are less likely to have potentially commercial activity—36.8 ± 10.1 percent, as compared to the entire sample.
- Parked domains are also more likely to be used for some kind of potentially commercial activity—90.3 ± 3.2 percent. This suggests that someone who encounters a domain that has little web/FTP content because it may be held for resale or other uses that do not involve web/FTP content will encounter some type of potentially commercial activity such as a banner or pay-per-click ad (see Appendix A, Table C.1).
- As might be expected, domains with little usable web/FTP content are the least likely to have potentially commercial activity—6.7 ± 2.4 percent. Occasionally, such a domain will contain a pay-per-click ad, some other type of ad or promotional content (see Appendix A, Table C.1).

Registrant's WHOIS Address Country/Region of the World

Exhibit 18: Country/Region of the World from Registrant's WHOIS Address Within Apparent Domain User Categories



Comparing the relative percentage of Registrant's WHOIS Address Country/Region of the World in **Exhibit 18** both within and across Apparent User Domain Type, we can observe the following associations between these two variables.

- Legal persons that use domain names are slightly less likely to have WHOIS addresses in the U.S.—54.9 ± 4.0 percent. This is statistically equivalent to the overall sample percentage of 56.9 percent.

    - The ranking of Registrant's WHOIS address county/region of the world for domains of legal person users is almost consistent[28] with overall sample ranking—that is U.S. (56.9 percent), Other Europe (11.2 percent), Other Asia/Pacific (9.0 percent), United Kingdom (5.0 percent; tied with China), China (5.0 percent), Germany (3.7 percent), Australia/New Zealand (3.4 percent), Canada (3.3 percent), Other (2.5 percent).[29]

- Domains with no usable web/FTP content (no usable online content) are slightly more likely to have WHOIS addresses in the U.S.—59.3 ± 5.2 percent, which is statistically equivalent to the overall sample percentage of 56.9 percent. The ranking of Registrant's WHOIS address county/region of the world for domains with no online content is also almost consistent with overall sample ranking

- Parked domains are the most likely to have WHOIS addresses in the U.S.—67.3 ± 5.1 percent, as compared to the entire sample. This suggests that someone who encounters a domain that has little web/FTP content because it may be held for resale or other uses that do not involve web/FTP content would be more likely to find that the domain name was registered in the U.S.

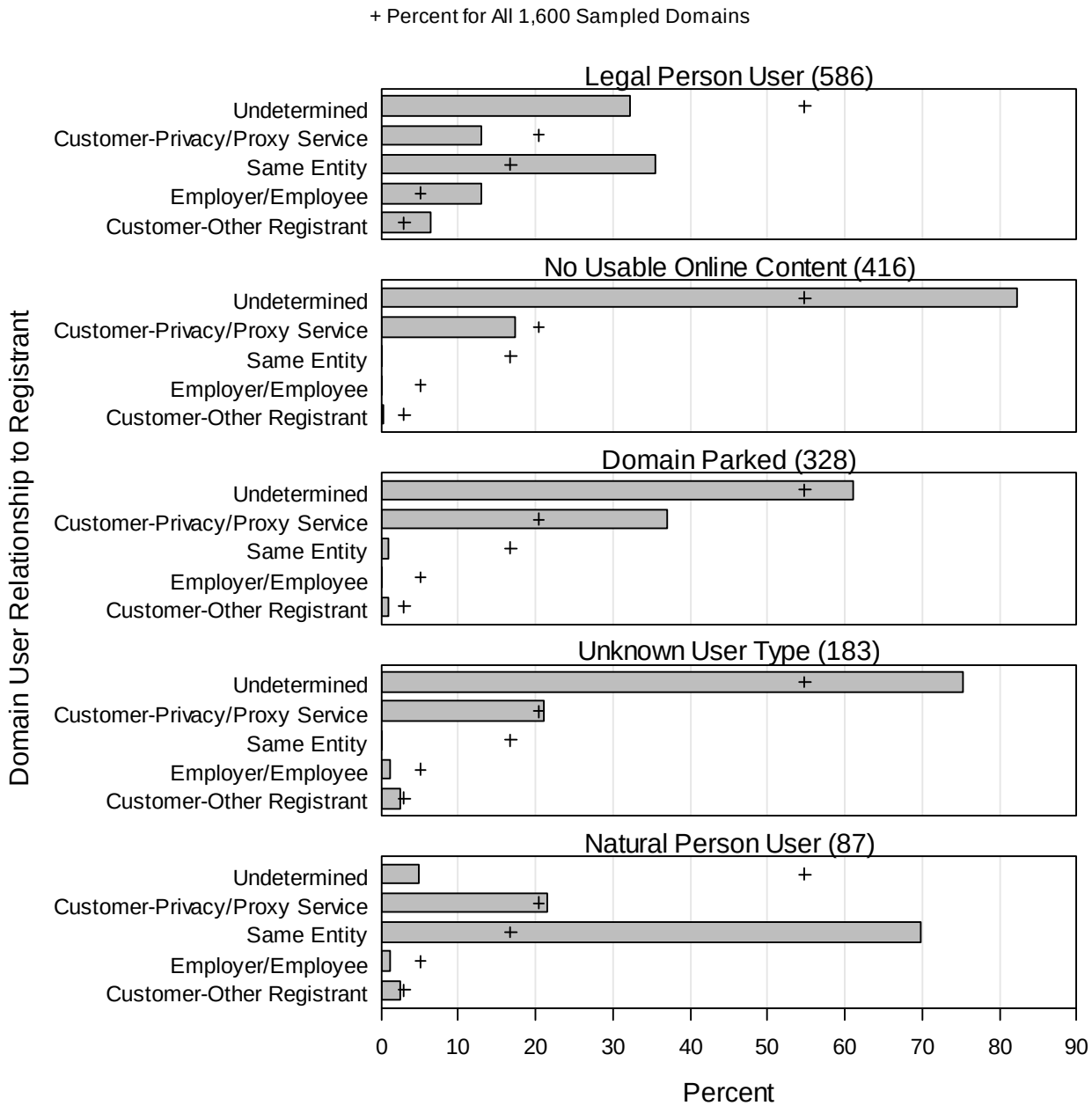
- The WHOIS address country relative percentages for natural person users are statistically equivalent to the entire sample percentages. For example, natural person users have 49.9 ± 10.4 percent of WHOIS addresses in the U.S. as compared to the entire sample's 56.9 percent. Germany has the appearance of being more likely in **Exhibit 18**; however, its relative percentage of 9.7 ± 6.2 percent is statistically equivalent to the entire sample's 3.7 percent WHOIS Germany addresses.

[28] There are some slight differences between the legal person user rankings versus the overall sample; however most of the differences are not statistically significant.

[29] Note that the overall sample percentages are similar to those shown in **Exhibit 3**. However, the percentages are relative to the 1,518 WHOIS records that did not have ambiguous or missing country address information.

Domain User Relationship to Registrant

Exhibit 19: Domain User Relationship to Registrant Within Apparent Domain User Categories



Comparing the relative percentage of Domain User Relationship to Registrant in **Exhibit 19** both within and across Apparent Domain User Type, we can observe that the overall sample ranking does not hold for legal person users or natural person users—that is, as shown in **Exhibit 6**, Undetermined (54.8 percent), Customer-Privacy/Proxy Service (20.4 percent), Same Entity (16.8 percent), Employer/Employee (5.1 percent), and Customer-Other Registrant (3.0 percent).

- Domain names used by legal persons were more likely to also be registered by that legal person—35.5 ± 3.9 percent, as compared to the entire sample's 16.8 percent. Domain names used by legal persons were also less likely to have the user/registrant relationship be a customer of a privacy/proxy service—13.0 percent ± 2.7 percent, as compared to the entire sample's 20.4 percent. For the other user/registrant relationship types, the relative percentages are similar to the overall sample percentages.

- Domain names used by natural persons were also more likely to be registered by that natural person—69.7 ± 9.6 percent, as compared to the entire sample. Domain names used by natural persons were also less likely to have undetermined user/registrant relationships—5.1 ± 4.6 percent, as compared to the entire sample's 54.8 percent. For the other user/registrant relationship types, the relative percentages are similar to the overall sample percentages.

The overall sample ranking does not hold within the other Apparent Domain User Types.

- Not surprisingly, when there is no usable web/FTP content, the user/registrant relationship is more likely to be undetermined—82.3 ± 3.7 percent, as compared to the entire sample's 54.8 percent.

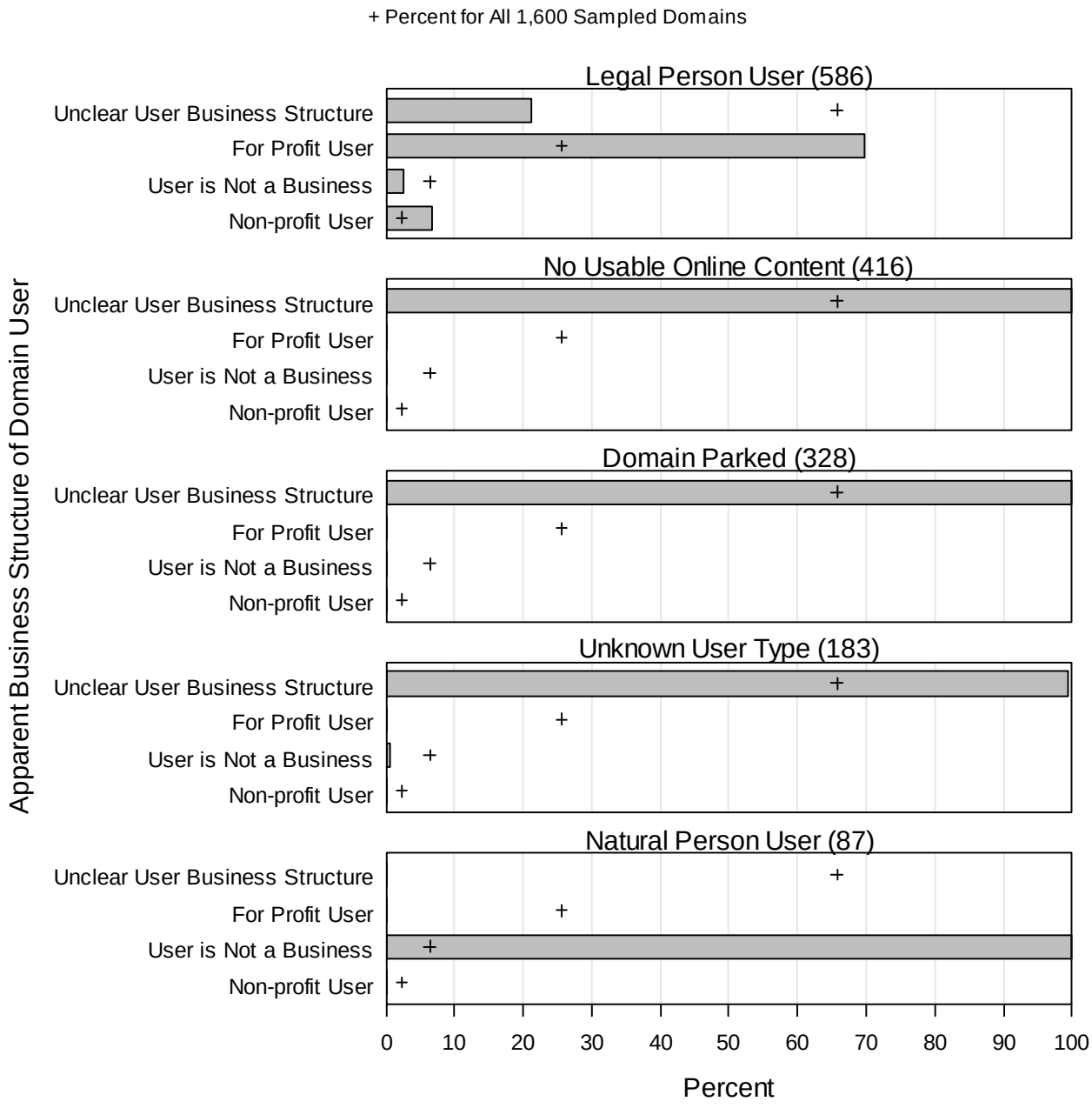
    - This relative percentage is not 100 percent because NORC's coding of this variable used the identity of the entity that presumably contracted a privacy service or other web development/consulting company to register the domains. In such cases, the registered name holder's identity was not shielded, and we could determine the relationship with the domain user.

- As might be expected, domain names used by entities with unknown user type were the most likely to have undermined user/registrant relationships—75.3 ± 6.3 percent, as compared to the entire sample's 54.8 percent.

- Parked domains are also likely to have undetermined user/registrant relationships—60.9 ± 5.3 percent. This is almost the same as for the entire sample. Parked domains are more likely to be customers of a privacy/proxy service—37.1 ± 5.2 percent, as compared with the entire sample's 20.4 percent.

Apparent Business Structure

Exhibit 20: Apparent Business Structure of Domain User Within Apparent Domain User Categories



Comparing the relative percentage of Apparent Business Structure in **Exhibit 20** both within and across Apparent Domain User Type, we can observe that except for legal person users, each domain user type consists of almost all (if not all) of one apparent business structure.

- For natural person users, the apparent business structure is never a business. This is by design—when coding apparent business structure, if the user was a natural person, then the business structure was coded as not a business.

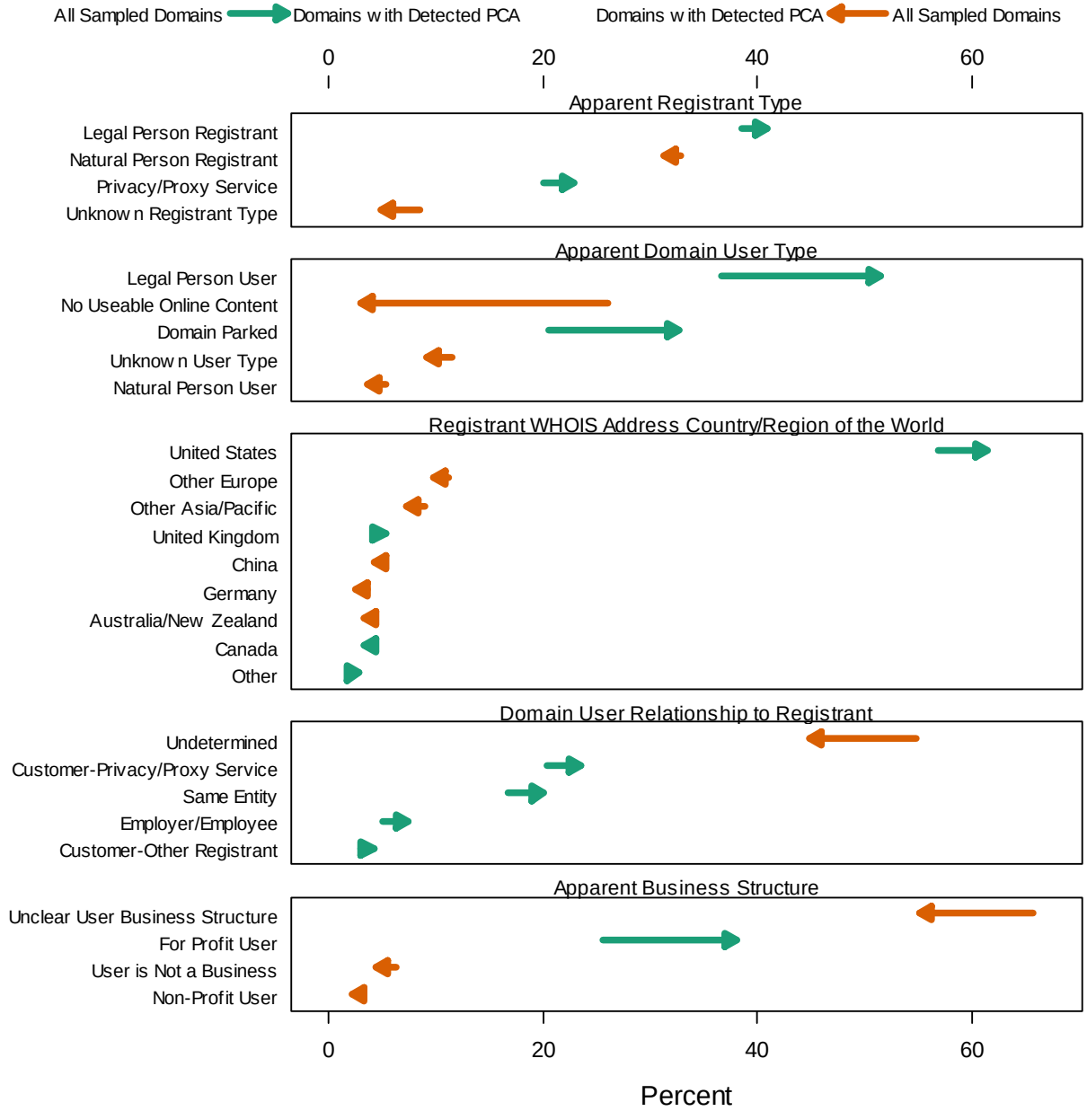
- Domain user categories "no online content" and "domain parked" always had an undetermined business structure because there was no information available to determine the business structure. We were unable to determine the business structure for almost all unknown users (99 percent). For a few unknown users (1 percent) we were able to classify the business structure as not a business.

- As might be expected, legal person users are more likely to be for-profit businesses—60.7 ± 3.7 percent, as compared to the entire sample's 25.6 percent (see **Exhibit 7**). Legal person users also are likely to be non-profit organizations—6.7 ± 2.0 percent, as compared to the entire sample's 2.3 percent.

3.3. Potentially Commercial Activity

In the previous sections, we examined the Potentially Commercial Activity relative to Apparent Registrant Type and Apparent Domain User Type. In this section, we will examine the relative percentage of domains in the various domain classification categories among 905 domains found to have potentially commercial activity as compare to the entire sample.

Exhibit 21: Potentially Commercial Activity Domains versus the Entire Sample



Comparing the percentage of each domain classification group in **Exhibit 21** both within each category (row of the graphic) and across the categories within each classification group (each panel of graphic); we

can observe the following associations between Potentially Commercial Activity and the classification categories.

- For both Apparent Registrant Type and Registrant WHOIS Address County/Region of the World differences between the relative percentage among domains with potentially commercial activity and the entire sample's percentage are small. Thus, knowing that a domain has potentially commercial activity does not provide any additional insight as to the registrant type or the WHOIS address of the registrant.

- Apparent Domain User Type differs between domains with potentially commercial activity and the entire sample.

    - As might be expected, legal person users are more likely among domains with potentially commercial activity—51.5 ± 3.3 percent, as compared to the entire sample's 36.6 percent.

    - Parked domains are also more likely among domains with potentially commercial activity—32.7 ± 1.9 percent, as compared to the entire sample's 20.5 percent.

    - Not surprisingly, domains with no usable web/FTP content (no online content) are less likely among domains with potentially commercial activity—3.0 ± 3.0 percent, as compared to the entire sample's 26.0 percent. As previously noted, some domains with little usable web/FTP content may have pay-per-click or other types of ads, but these were seldom found.

- Domains for which the user/registrant relationship could not be determined are less likely among domains with potentially commercial activity—44.8 ± 3.2 percent, as compared to the entire sample's 54.8 percent.

- Apparent Business Structure has difference between domains with potentially commercial activity and the entire sample.

    - Not surprisingly, domains with users that are for-profit businesses are more likely among domains with potentially commercial activity—38.1 ± 3.2 percent, as compared to the entire sample's 25.6 percent.

- Domains for which the business structure was unclear are less likely among domains with potentially commercial activity—55.2 ± 3.2 percent, as compared to the entire sample's 65.7 percent.

Exhibit 21 also provides information to answer the GAC question: ***What is the relative percentage of Privacy/Proxy use among domains with commercial use?***

In our random sample of 1,600 domains, 905 were found to have potentially commercial activity. Of these domains, 22.9 percent (± 2.7 percent) were registered using a privacy or proxy service. This is not statistically different from the 20 percent of domains registered using a privacy or proxy service in the entire sample.

4: Lessons Learned

NORC learned some valuable lessons for conducting a study of this nature. Below we summarize what we feel are some of the more important lessons related to data collection and data coding. We recommend that future ICANN studies of this nature use these lessons as a starting point.

Data Collection

As noted in section 2.2, due to the fluid nature of domain content, the amount of lapsed time between sample domain content extractions (including WHOIS lookup) introduces a potential for data stagnation error. NORC attempted to reduce the amount of time between content source extractions for a given domain to potentially minimize this error by conducting three content source extractions simultaneously. We believe that this provided the best possible "snapshot" of the sampled domains at a given point in time. We developed a Python-based application, the NORC-BOT, to accomplish this.

NORC-BOT distributed the tasks associated with extracting content sources across three threads that ran in parallel:

- WHOIS data: this thread made API calls on http://whoisxmlapi.com/, requesting the WHOIS information in XML format. [30]
- Publically accessible HTTP/HTTPS/FTP files: this thread downloaded a fixed amount of publically accessible files on a given domain name and the www subdomain using HTTP, HTTPS, and FTP.[31] Setting a download quota (100MB) was necessary to ensure that extremely large sites hosting GBs of content were not indexed. To collect this content the GNU Wget tool (http://www.gnu.org/software/wget/) was used.
- Response codes from DNS BlackLists (DNSBL) for the given domain. For each domain, threading was also introduced at the task level, specifically for the DNSBL task, to ensure that processing of the domains was completed in a timely manner.

In what follows, we summarize important aspects of these threads. This knowledge may benefit others that may attempt similar exercises in the future.

[30] whoisxmlapi is a third party service requiring purchase of API calls. To complete this project NORC purchased 5000 queries.

[31] Publically accessible indicates the NORC-BOT will not attempt to download password protected content and the NORC-BOT will respect the Robot Exclusion Standard (RES) specification listed in the sites robots.txt file.

WHOIS Thread

To collect WHOIS data, NORC used *WhoisxAPI*, a web-based service that returns WHOIS data for a domain through an HTTP request (http://whoisxmlapi.com/). The service returns the listed WHOIS information in a machine-parsable format, either json or xml. *WhoisAPI* proved to be an effective service; however, in our initial runs, we encountered timeout errors (resulting from the *WhoisAPI* service not sending a back reply to the NORC-BOT's HTTP GET request) and internal whoisxmlapi server errors. Through discussions with a representative at whoisxmlapi, NORC learned that a non-documented parameter, hard_refresh=1, could be supplied in the GET request to invoke the service to supply a non-cached WHOIS record. Testing found that this parameter significantly decreased the number of errors WHOIS extraction errors. Using this parameter, however, costs five queries instead of one for each domain. In all, NORC purchased 5000 queries in order to have sufficient API call resources for completing data extraction.

The service proved very effective, provided WHOIS data was in a format *WhoisAPI* expected.[32] Where WHOIS data was not listed in readily separable formats, *WhoisAPI* returned the data in "na" unparsed node, and NORC had to employ manual review and data cleaning to extract usable data from this unparsed field. Any missing or incorrectly parsed WHOIS registrant information was initially searched for in the unparsed field. If it could not be found, we used the WHOIS information extracted by ICANN staff to fill in gaps. The ICANN data was used if all fields were missing, if there were gaps in WHOIS returned by the API, we checked the ICANN WHOIS dataset and, if the records were clearly related, we filled in the missing WHOIS information from the ICANN dataset.

HTTP/HTTPS/FTP Thread

Regarding downloaded Internet content, NORC noted two potential exceptions to our largely successful procedures. Some domain names may not have returned web/FTP content because the domains were being used for other purposes, such as for mail services. Additionally, some of the web pages NORC downloaded were local copies, and if the content had dynamic content or Flash content, we did not retrieve the full set of information. Any links embedded within the Flash app could not be followed by WGET (the program used by NORC-BOT); as a result, the extraction of such sites may be incomplete. Second, all content embedded in the Flash application could not be rendered as HTML content, precluding further analysis by coders.

[32] Different registrars present WHOIS data in different ways, some of which may have been unfamiliar to the *WhoisAPI* service, and therefore could not be parsed. Additionally, registrants do not always populate all WHOIS fields, or misunderstand what to put in each field, contributing to parsing errors.

DNSBL Thread

In an effort to determine allegedly illegal or harmful activities present in our sample, DNSBL lists were scanned for each sampled domain. The DNSBL strategy was to obtain all the A RECORDS associated with the sampled domain. For each A RECORD, the returned IP address was checked against a series of DNSBLs. After the collection of DNSBL data, NORC aggregated the specific response code data for all the DNSBLs where a match had been detected. For all DNSBLs that returned a match, NORC conducted an in depth review of the DNSBL evaluating the DNSBL listing methodology, the apparent status of how actively the list is being maintained, and the documentation on the set of response codes returned by the list. NORC used this information to determine which DNSBLs were most likely to have provided accurate information with which to identify allegedly illegal or harmful activity. Ultimately, this review produced a document containing information on the DNSBL service and its response codes. We learned that the reliability of the response codes was suspect. Therefore, each response code was given an accuracy score determined by the factors listed above. After the response code information was gathered for all the DNSBL-response code combinations, the DNSBL matches for each domain were cross-referenced against this table. All matches with an accuracy score below 3 were removed. The remaining DNSBLs were recorded into a series of variables: one for each DNSBL-response code combination relevant to the study.

Data Coding

While data collection in this project posed many technical challenges, data coding provided more subjective challenges due to the inherent ambiguity of internet data. NORC followed an iterative process to develop precise yet inclusive code frames for the variables of interest to this project, beginning with standard classification of many variables and then adjusting these code frames, or creating supplemental variables, to accommodate the complexity presented by the data. The attempt to impose standard codes on a huge variety of unique websites also revealed the fuzziness of some prevailing concepts used in studying Internet activity.

One example of this fuzziness is the issue of domain parking versus domain reselling. Domain reselling—the act of purchasing a potentially attractive domain name with the intent simply to resell it to someone with a functional interest in that specific domain name—is one among several potential intentions behind domain parking. In other cases, domain registrants may have purchased the domain with the intent to use it themselves, but have not yet created an actual webpage or posted any content to it. Domains may also be parked simply to generate income for registrants by picking up traffic on commonly misspelled websites addresses and exposing web browsers to banner and pay-per-click ads. Although the registrant's intent cannot be known for certain, when encountering these cases, coders attempted to discern whether there was a resale intent by executing reverse WHOIS queries at reversewhois.domaintools.com to

determine if the domain registrant was a multiple domain name holder (with "multiple" set at a conservative 50 domains). Possible intent for future use by the registrant, conversely, was inferred from a judged correspondence between the domain name and other WHOIS information (such as registrant organization name).

Several variables proved to be particularly vexing because of the obscurity or ambiguity of the type of information they recorded. These variables include of Domain User Relationship to Registrant, Apparent Business Structure of Domain User, and Apparent Business Function (ultimately discarded due to coding difficulties). Of these, the Relationship variable was perhaps the most difficult to determine. Although the code frame included many coding options, only some were readily discernible—registrant is the same entity as the user, registrant is a proxy service—it was often impossible to determine if a registrant was a hosting provider, employee, or client of the domain user. This difficulty is reflected in the prevalence of "undetermined" in this variable. Since NORC's data collection was rather comprehensive on this project, it appears the only way to gain more fidelity in this variable would be for registrants to supply this relationship as part of the WHOIS information.

We also developed code frames for the business structure and function of the domain users who appeared to be legal persons. Basing our initial code frame on standard business classifications, we attempted to distinguish between corporations and smaller entities such as partnerships and sole proprietorships in the business structure variable. The same process in the business function variable yielded fourteen codes; among these were traditional functions such as Enterprise, Retail, Non-Profit, Consultancy, as well as newer digital functions, such as Utility and Domain Parking. Initial coding attempts revealed the difficulty of making clear determinations within these variables; there were 943 domains (58.9 percent) whose business structures and functions could not be discerned by our coding team. We found that while some designations, such as Corporation (in structure) and Enterprise (in function) were often prominently stated or easily inferred from web content, many other determinations were either impossible to make or admitted ambiguity with overlapping possible designations. For example, while it was often clear that a website promoted a small business, it would not be clear whether this business was a partnership or a sole proprietorship. In the business function variable, many businesses often fit within multiple categories and required rather fine distinctions to be made. Keyword searches were used to help with this effort, but there were not enough keywords identified to make an automated process reliable.

We addressed these difficulties by completing two rounds of coding for each sampled domain, with extensive training sessions before each to explicate the common distinctions we expected our coders to make. In both rounds, coders were given a standard code frame to reference in their coding, while also having the latitude to note unique circumstances for each domain. Coders analyzed not only the

downloaded web content for references to business structure and indications of function, but also employed third party services such as Accurint and LinkedIn to provide supplemental information or corroboration for codes. Online translators (including Google's translation function) were used to decipher foreign language pages. After this process, analysts performed adjudication of these two sets of codes in order to reconcile discrepancies into a set of codes with more uniformly applied rules, while also taking into account the special circumstances noted by coders. Adjudicators also developed a set of generic structure and function codes to consolidate the myriad designations into more abstract categories for analysis (thus, codes such Partnership and Sole Proprietorship in the Apparent Business Structure variable were consolidated into the "Small Business" code in a generic Business Structure variable). Throughout the process, we emphasized that indecipherable cases should be coded as one of the various "Unknown" codes to maintain the high-quality nature of the data.

5: Conclusions and Recommendations

NORC set out to perform a study of the top five gTLDs in order to better understand registrant identification issues. We have gathered a set of data that is useful for its intended purpose—an exploratory study of registrant and domain user identifications, characteristics and the types of domain use activities. In particular, we focus on analyses related to the following three questions.

1) What differences exist in how domains are actually used for domains registered by natural persons versus domains registered by legal persons versus domains registered via proxy?

2) What differences exist between how domains users that are natural persons identify themselves, versus how domain users that are legal persons identify themselves?

3) What differences exist in how domains with any type of potentially commercial activity are identified in WHOIS versus domains with no observed potentially commercial activity?

In many cases, classification of the characteristics and activities were difficult to discern and often had to be coded as "unknown." However, a large enough number of domains were able to be coded so that important relationships were uncovered. The ICANN community will find this information useful for fact-based WHOIS policy discussions.

In terms of answering the three questions listed above, the data reveal the following highlights.

- Differences in how domains are used based on domains registrant type (see section 3.1 for more results).
    - Apparent legal person registrants are more likely to be legal person domain users (52.2 ± 3.9), and their domains are likely to have potentially commercial activity (59.9 ± 3.9 percent).
    - Domain names registered using a privacy/proxy service were the most likely to be used for some kind of potentially commercial activity (64.6 ± 5.2 percent), and they were more likely to be parked (30.7 ± 5.0 percent), as compared to all registrant types (20.5 percent).
    - Domain names registered by natural persons are more likely to be used by natural persons (10.4 ± 2.6 percent), as compared to all registrant types (5.4 percent).

- Differences in how domains users identify themselves based on domains registrant type (see section 3.2 for more results).
    - Domain names used by legal persons were more likely to be registered by legal persons (55.1 ± 4.0 percent), as compared to all domain users (38.6 percent), and were more likely to be used for some kind of potentially commercial activity (79.8 ± 3.2 percent), as compared to all domain users (56.6 percent).
    - Domain names used by natural persons were more likely to be registered by natural persons—60.4 ± 10.2 percent, as compared to all domain users 32.8 percent, and were less likely to have potentially commercial activity (36.8 ± 10.1 percent), as compared to all domain users (56.6 percent).
- Differences in how domains with potentially commercial activity are identified in WHOIS based on registrant type (see section 3.3 for more results).
    - Legal person registrants make up 40.9 percent (± 3.2 percent) of domains with potentially commercial activity.
    - Natural person registrants make up 31.3 percent (± 3.0 percent) of domains with potentially commercial activity.
    - Domains registered using a privacy/proxy service make up 22.9 percent (± 2.7 percent) of domains with potentially commercial activity.

NORC recommends that ICANN continue to examine WHOIS registrant identification issues and expand upon the work we have done. In particular, a review of the rules used to code variables should be done to determine the sensitivity of the results to changes in the rules. For example, in Appendix A, section C, NORC looked at changes in the number of domains having potentially commercial activity if pay-per-click ads are not included in the definition. While potentially commercial activity percentages are reduced, the relative ranking of registrant and user types are similar compared to using pay-per-click ads in the definition of potentially commercial activity. Thus, many of the observations related to potentially commercial activity are not sensitive to defining the variable with or without pay-per-click ads.

NORC also recommends that ICANN study alternative ways to code the data in order to reduce the number of "unknowns." A variable such as "Apparent Business Structure" contains a large percentage of domains used by entities with unclear business structure. NORC tried to find better ways to code this variable, including the use of keywords searches, but the variable still remained hard to code. This

variable was suggested by NORC during the study design in the hopes that it might provide a way to understand the relationship between registrants and users. Perhaps business structure is an ill-defined concept given the vast array of businesses around the world. If so, ICANN should seek to find other concepts that might better explain the registrant/user relationship.

Overall, the WHOIS Registrant Identification Study was successful despite the coding problems encountered. The collected data reveal relationships that the ICANN community wants to understand, and the data can be used to guide future policy.

ICANN WHOIS REGISTRANT IDENTIFICATION PROJECT

Appendix A: Exploratory Analysis Report

PRESENTED TO:
ICANN

PRESENTED BY:
NORC at the University of Chicago

MAY 23, 2013

NORC at the UNIVERSITY of CHICAGO

Introduction

NORC has been contracted by the Internet Corporation for Assigned Names and Numbers (ICANN) to conduct the **WHOIS Registrant Identification Study**; an exploratory study to classify domains into a variety of categories such as registrant type, domain user type, and commercial activity.

In creating the data we have collected, we have kept in mind the three focus questions of this project:

1) What differences exist between how domains users that are natural persons identify themselves, versus how domain users that are legal persons identify themselves?
2) What differences exist in how domains are actually used for domains registered by natural persons versus domains registered by legal persons versus domains registered via proxy?
3) What differences exist in how domains with any type of potentially commercial activity are identified in WHOIS versus domains with no observed potentially commercial activity?

We start the report with some background on these three questions, including how we recoded variables in the dataset for analysis. Our analysis is organized by how these three questions are answered for different subject variables. The first three analysis sections are the variables from which we have formed the three questions. Here are the subject variables for which we have analysis sections:

A. Apparent domain user type
B. Apparent registrant type
C. Potentially commercial activity variables
D. Business Structure of Domain User
E. Domain name extension (gTLD)
F. Registrant country/region of the world
G. Relationship of domain user to registrant
H. Other coded behavior variables
I. Blacklist variables
J. Whitelist variables

Our key tool for our analyses has been the chi-square test of independence.[1] Since this is an exploratory data analysis, we mainly interpret the frequencies rather than create more complex analysis such as

[1] A chi-square test of independence is a statistical test for assessing whether two categorical variables are independent (not associated). The null hypothesis of the test is that the two categorical variables are independent. If the observed chi-square test statistic, which is based on the difference between observed and expected cross-classified frequencies, is unusually large assuming the null hypothesis of independence is true, then we conclude

building regression models. Follow-up analyses can be done with the clearer focus that will come out of this project.

One important note is that all of our analyses except the one-way frequencies of variables are weighted. In a representative sample of 1,600 domains, we would have studied only 98 *.info and 26 *.biz domains, but we set sample sizes of 100 for each. We did this so that we could have a sufficient number of *.info and *.biz domains for analysis. This results in a slight undersampling of *.com, *.net, and *.org domains, and oversampled *.info domains and especially oversampled *.biz domains. So we applied weights to each gTLD as shown in Table 1.

Table 1: Weighting by gTLD for the Registrant ID Study Domain Sample

gTLD	Global Proportion	Sample Size	Sample Proportion	Weight = Global/Sample Proportion	Sum of Weights = Sample Size *Weight
*.com	74.3%	1,128	70.5%	1.0534	1188.2
*.net	10.7%	165	10.3%	1.0412	171.8
*.org	7.2%	107	6.7%	1.0813	115.7
*.info	6.1%	100	6.3%	0.9830	98.3
*.biz	1.6%	100	6.3%	0.2600	26.0
TOTAL	**100.0%**	**1,600**	**100.0%**		**1,600.0**

that the two categorical variables are associated (dependent upon one another). If the p-value—the probability, under the null hypothesis, of observing a test statistic value greater than or equal to the one obtained from the sample, is small, then the observed test statistic is considered unusually large. If you want at least 95 percent confidence for statistical test results, p-values less than 0.05 (5 percent) are considered too small. In this sense, we state that the chi-square test results are statistically significant.

The Three Questions

Apparent Domain User Type: Legal and Natural Persons

For each of the 1,600 domain names, we tried to determine if the domain user could be considered a legal person or a natural person. Table 2 shows that for most domain names, we could not make such a determination because almost half the domains were parked domains or had no online content at all. Only 11.5 percent of the domains had content, but had an unknown apparent domain user type. To code apparent domain user type, NORC staff reviewed all of the downloaded domain content for each domain during phase I of the Domain User variable coding. The overall procedure can be summarized as follows.

First, the downloaded web content was accessed to determine if the downloaded web content contained any usable data to conduct manual coding. If the data did not contain enough usable information, it was considered having No Usable Content and the Domain User variables relying on web content for coding were coded to their corresponding unknown codes. An example of this scenario is if the downloaded content consisted of a single webpage which only contained the following HTML data: <html><body> Under Construction </body></html>.

For the domains with usable data, we evaluated the downloaded content to determine if it consisted solely of common domain parking content. For example, if the full set of downloaded content consisted of a single landing page and this landing page only contained HTML content consistent with GoDaddy parking services, the apparent domain user type was coded as Unknown – Domain Parked. In some cases, it was not clear whether we should classify a domain as Domain Parked or No Online Content. Some of the No Online Content domains actually have a little content, and sometimes even some potentially commercial activity. For example, a site could have a simple index.html with an Under Construction page with a simple banner ad. There were not enough such sites to create a separate "Little Online Content" category.

All the domains which were not coded by the two procedures listed above were evaluated on a case-by-case basis to determine the phase I Domain User variables. The Apparent Domain User type was coded as a Natural Person when the Domain User was clearly a real living individual or small group of individuals and not a virtual entity such as a corporation or non-profit entity of any other named entity that is not a real living person. All other entities were coded as Legal Persons or Unknown.

To ensure that the data was accurately coded, each case underwent multiple rounds of manual coding by independent coders. The results of these multiple rounds of coding were adjudicated and all differences

detected during adjudication were collaboratively reviewed by a supervisory team to make a final determination of the Domain User variables.

Table 2: Apparent Domain User Type

Type	Frequency	Percent
Natural Person	87	5.4
Legal Person	586	36.6
Domain Parked	328	20.5
No Online Content	416	26.0
Unknown	183	11.5

A finer categorization of Natural Person was done to separate the variable into individuals versus small groups of related individuals, for example, a family. We found that of the 87 Natural Persons shown in Table 2, 78 are individuals and nine are small groups. Further analysis of the group of nine domains would not provide statistically meaningful results, so we will not split the Natural Person category in subsequent analyses. Analyses will only compare the three generic entity types: legal persons, natural persons, and unknown.

Registrants: Natural and Legal Persons and use of Privacy/Proxy Services

Apparent registrant type was coded as to whether we could place the registrant into categories defined in ICANN's *Revised Terms of Reference for WHOIS Registrant Identification Studies* (http://gnso.icann.org/issues/whois/tor-whois-registrant-id-studies-20may11-en.pdf). Initially, only WHOIS information and independent searches of public databases were considered in the classification. For example, we searched known lists of privacy and proxy providers to place sampled domains into these categories, and reverse WHOIS email counts were used to help determine multiple domain name holders. Manual coding was used to code the remainder of the domains where Apparent Registrant Type could not be classified using automated means. The Apparent Registrant Type was coded during phase I of the Domain User Coding process. This manual coding process consisted of a concise set of rules to arrive at Apparent Registrant Type. The manually coded cases underwent the same quality control process consisting of multiple rounds of independent coding and an adjudication process. While investigating the domain user, the coder may have gained insights on the registrant of the domain, such as situations where the domain user is the same as the registrant. Thus, additional information was used to

correct initial categorizations or add granularity to the process. Table 3 is a summary of the final coding outcomes for Apparent Registrant Type:

Table 3: Apparent Registrant Type Summary

Apparent Type	Frequency	Percent
Registrant Name appears to be a natural person; no organization is named	447	27.9
Registrant Organization is specified; registrant name is also specified – registrant name or organization contains legal person	320	20.0
Registrant Organization appears to be a Proxy registration service	300	18.8
Registrant Organization is specified and appears to be a legal person; no registrant name is specified	183	11.4
Registrant Name and Organization are completely missing	93	5.8
Registrant Organization is specified; registrant name is also specified – both appear to be a natural person	73	4.6
Registrant Organization appears to be a multiple domain name holder	62	3.9
Registrant Name appears to be a legal person; no organization is named	52	3.3
Registrant Name and Organization look to be patently false	25	1.6
Registrant Organization appears to be a Privacy registration service	20	1.3
No Registrant Name or Organization available because Pending Reactivation or Deletion	11	0.7
Unable to classify / requires additional review	7	0.4
Registrant Organization is specified and appears to be a natural person; no registrant name is specified	5	0.3
Registrant Name and Organization are incomplete	2	0.1

With respect to the questions that are the key focus of this study, domains that are registered using Privacy or Proxy services are of particular interest. As shown in Table 3, there are 300 proxy-registered domains, but only 20 privacy-registered domains. With such a small category size, further analysis that attempts to cross-classify the privacy group with subject variables, such as commercial activities, would not be meaningful. Therefore, our analyses combine privacy and proxy registered domains together, though it is almost a comparison between proxy and non-proxy registered domains.

In order to simplify analyses of Apparent Registrant Type, we collapse the categories in Table 3 to the following four revised categories:

- Registrant appears to be a Legal Person – domains with WHOIS data which appears to identify a legal person as the Registrant (includes multiple domain holders, but not Privacy/Proxy registered domains)
- Registrant appears to be a Natural Person – domains with WHOIS data which appears to identify a natural person as the Registrant
- Registrant appears to reference a Privacy/Proxy Service – domains with WHOIS data which appears to identify a Privacy/Proxy service
- Unknown – domains with WHOIS data which could not be classified (includes data completely missing, patently false or incomplete WHOIS, and domains pending reactivation or deletion)

In what follows, the term Apparent Registrant Type refers to these revised categories. Table 4 is a summary of Apparent Registrant Type revised.

Table 4: Apparent Registrant Type Summary (Revised)

Apparent Type	Frequency	Percent
Registrant appears to be a Legal Person	617	38.6
Registrant appears to be a Natural Person	525	32.8
Registrant appears to use a Privacy/Proxy Service	320	20.0
Unknown	138	8.6

Potentially Commercial Activity

There are several variables related to potentially commercial activity in the domain content section of the dataset. These variables measured whether there was any apparent activity that might be considered commercial in some countries: whether there were membership dues for online content or offline content, whether there was promotional content offline or online, whether there were banner ads and whether these banner ads were for the hosting provider or registrar, and whether there were only pay-per-click ads and whether these pay-per-click ads were for the hosting provider or registrar. We created a variable measuring Potentially Commercial activity in any of these variables. All of these variables are binary, so these tables only present the percentage of domains in each subgroup with each of these characteristics. Table 5 shows the overall percentage for each binary variable.

Table 5: Summary of Potentially Commercial Activity Variables

Commercial Activity Variable	No	Yes	Percent Yes
E-Commerce	1489	111	6.9
Membership (Online Content)	1572	28	1.8
Membership (Offline Content)	1544	56	3.5
Promotional Content (Offline)	1305	295	18.4
Promotional Content (Online)	1507	93	5.8
Host Promotional Content (Online)	1461	139	8.7
Third Party Banner Ads	1496	104	6.5
Host Banner Ads	1398	202	12.6
Pay-Per-Click Ads	1131	469	29.3
Host Pay-Per-Click Ads	1539	61	3.8
Any Potentially Commercial Activity	695	905	56.6
Excluding Pay-Per-Click Ads	883	717	44.8

A further explanation of coding these variables is described below:

E-Commerce

This classification variable allows for e-commerce activities to be noted for any site, even if the site is not primarily an "e-commerce" website. For instance, ESPN.com, while classified as an "informational" website, would here receive a value of "1" (true) since ESPN.com provides pages where website readers can purchase goods from ESPN.com.

Membership (Online Content)

Membership fees will typically require a user name and password for logging in to view privileged online content.

However, many websites will ask for users to create user names without charging a membership fee; the user name creation allows these websites to gather information on its users and communicate better with these users, thereby increasing traffic to the website. These types of membership are NOT marked as having commercial online membership.

To determine if member logins first require the payment of membership fees, we went to the login page of the website to see if membership is offered for a price. Sometimes, fees are not immediately apparent; for instance, the New York Times allows specific computers to access New York Times online content

ten times per month before requiring a membership fee-based login to access its content. Because of mechanisms like this, we had to carefully assess the membership requirements of the site.

Membership (Offline Content)

As opposed to online membership, offline membership refers to fees paid through the website for goods or services provided offline. For example, a gym may offer a portal through which gym members pay their monthly membership fees so that they may continue to use the physical gym.

Promotional Content (Offline)

Promotional content encourages website visitors to purchase goods or services of the website owner, either in a physical location or through some other vendor, instead of through the website itself. Promotional content is distinct from e-commerce activity because the commercial activity is merely being promoted, but cannot be transacted, on the website in question.

An example of a website with promotional content would be a small bookstore website that advertises its latest book arrivals on its website, but which does not have a web portal through which these books can be purchased online; a customer must go to the physical location of the bookstore in order to purchase the books.

Promotional Content (Online)

If a website is promoting their goods but these goods are sold on an online retailer site like Amazon or E-bay, then this is an example of PROMO-ON.

Host Promotional Content (Online)

Same as promotional content described above, but there is evidence that the promotional content was placed on the website by the hosting provider.

Third Party Banner Ads

Banner ads are graphics on websites which advertise goods or services and which act as links to pages where these goods or services can be purchased online. The placement of these ads on third party sites allows the domain users of these sites to earn revenue from the companies placing the ads. Note that these banner ads are shown regardless of the type of site visitor or the type of content they are viewing. This is opposed to pay-per-click ads, which generally appear in response to specific queries by site visitors.

Discerning whether the domain user or the hosting provider placed the banner ads on the website can be difficult. Generally, websites that appear to be administered or designed by the domain user will be more

likely to have ads that were placed by the domain user (since the domain user is exercising a large amount of control over the domain).

Conversely, if the site template is provided by the host, or if the hosting appears to be free, it is likely that the host is placing ads on the site (this would be part of the agreement for free hosting).

This variable asks simply whether banner ads are present on the site.

Third Party Banner Ads Host

A determination of whether banner ads placed by hosting providers are present on a website, following the distinctions from the preceding variable description.

Pay-Per-Click Ads

Pay-per-click ads, unlike banner ads, appear in response to site visitor queries or the type of content the visitors view. This occurs because pay-per-click ads generate revenue for domain users or hosting providers based on “performance” (number of clicks) rather than “impressions” (number of views).

Some websites appear to exist only to generate these types of ads; this variable tracks this type of website.

Host Pay-Per-Click Ads

Same as Pay-Per-Click Ads but there is evidence that the hosting provider placed the ads on the website.

Analyzing the table as a whole, since the sum of the individual Yes variables is 1,558, there are many domains with more than one type of potentially commercial activity (average of 1.72 activities for those with at least one). The most common activity in Table 5 is pay-per-click ads, which might not be considered to be potentially commercial activity by some. Therefore, we also calculated a version of the potentially commercial activity excluding domains with only pay-per-click ads. This excluded 188 domains, lowering the estimate to 44.8 percent.

A. Apparent Domain User Type

Apparent Registrant Type

Table A.1: Apparent Domain User Type by Apparent Registrant Type Weighted Cross-classified Frequency Counts

Apparent Domain User Type	Apparent Registrant Type									
	Natural Person		Legal Person		Privacy/Proxy		Unknown		Total	Percent
Natural Person	53.7	60.4	10.8	12.1	20.3	22.8	4.2	4.7	88.9	5.6
Legal Person	147.6	25.1	324.5	55.1	89.0	15.1	27.3	4.6	588.4	36.8
Domain Parked	116.5	35.4	95.7	29.1	98.8	30.0	18.2	5.5	329.2	20.6
No Online Content	128.9	31.2	130.5	31.6	75.0	18.2	78.1	18.9	412.5	25.8
Unknown Type	68.6	37.9	60.4	33.4	39.2	21.6	12.9	7.1	181.1	11.3
Total Percent	515.3	32.2	621.8	38.9	322.3	20.1	140.7	8.8	1600	100

There is a strong relationship between apparent domain user type and apparent registrant type, with a p-value for the relationship of less than .0001. Overall, 32.2 percent of registrants are apparently natural persons, but for apparent natural person domain users, this percentage is 60.4. Overall, 38.9 percent of registrants are apparently legal persons, but for apparent legal person domain users, this percentage is 55.1 percent. Only 12.1 percent of the apparently natural persons use domains registered by apparently legal persons. Overall, 20.1 percent of the domains are apparently registered using a privacy/proxy service. This percentage is highest for the domain parked domains (30.0 percent) and lowest for legal person domain users (15.1 percent). Overall, 8.8 percent of the domains have unknown registrant types, but this percentage is 18.9 percent for domains with no online content. (Note: Online content was not used to determine registrant type.)

Potentially Commercial Activity

Table A.2: Apparent Domain User Type by Potentially Commercial Activity Weighted Cross-classified Frequency Counts

Apparent Domain User Type	Potentially Commercial Activity					
	Not Detected		Detected		Total	Percent
Natural Person	56.2	63.2	32.7	36.8	88.9	5.6
Legal Person	119.1	20.2	469.3	79.8	588.4	36.8
Domain Parked	31.8	9.7	297.4	90.3	329.2	20.6
No Online Content	384.9	93.3	27.6	6.7	412.5	25.8
Unknown Type	97.4	53.8	83.7	46.2	181.1	11.3
Total Percent	689.4	43.1	910.6	56.9	1600	100

There is a strong relationship between apparent domain user type and Potentially Commercial activity, with a p-value for the relationship of less than .0001. Overall, 56.9 percent of domains show Potentially Commercial activity, but this is highest for domain parked domains (90.3 percent) and apparently legal person domain users (79.8 percent). Potentially Commercial activity was detected for only 6.7 percent of the domains with no online content (page 4 gives a fuller explanation of an Under Construction page with no online content other than a simple banner ad).

B. Apparent Registrant Type

Apparent Domain User Type

Table B.1: Apparent Registrant Type by Apparent Domain User Type
Weighted Cross-classified Frequency Counts

Apparent Registrant Type	Apparent Domain User Type: Natural Person		Legal Person		Domain Parked		No Online Content		Unknown Type		Total	Percent
Natural Person	53.7	10.4	147.6	28.6	116.5	22.6	128.9	25.0	68.6	13.3	515.3	32.2
Legal Person	10.8	1.7	324.5	52.2	95.7	15.4	130.5	21.0	60.4	9.7	621.8	38.9
Privacy/Proxy	20.3	6.3	89.0	27.6	98.8	30.6	75.0	23.3	39.2	12.2	322.3	20.1
Unknown	4.2	3.0	27.3	19.4	18.2	12.9	78.1	55.5	12.9	9.2	140.7	8.8
Total Percent	88.9	5.6	588.4	36.8	329.2	20.6	412.5	25.8	181.1	11.3	1600	100

As we discussed for Table A.1, there is a strong relationship between apparent registrant type and apparent domain user type, with a p-value for the relationship of less than .0001. This table is just Table A.1 with the rows and columns reversed. Only 5.6 percent of the domain users are apparently natural persons, but this percentage is almost doubled (10.4 percent) for registrants that are apparently natural persons. The lowest percentage of domain users that are apparently natural persons are for registrants than are apparently legal persons (1.7 percent). Overall, 36.8 percent of the domain users are apparently legal persons, but this percentage is 52.2 percent for registrants that are apparently legal persons. Overall, 20.6 percent of the domains were parked (preventing further user classification), and this percentage is highest for privacy/proxy registered domains (30.6 percent) and lower for registrants who are apparently legal persons (15.4 percent). Overall, 25.8 percent of the domains had no online content, but this percentage is 55.5 percent for unknown registrant types. Roughly ten percent of the domain users have an unknown type, regardless of the apparent registrant type.

Potentially Commercial Activity

Table B.2: Apparent Registrant Type by Potentially Commercial Activity Weighted Cross-classified Frequency Counts

Apparent Registrant Type	Potentially Commercial Activity					
	Not Detected		Detected		Total	Percent
Natural Person	229.6	44.6	285.6	55.4	515.3	32.2
Legal Person	249.5	40.1	372.3	59.9	621.8	38.9
Privacy/Proxy	114.0	35.4	208.3	64.6	322.3	20.1
Unknown	96.2	68.4	44.4	31.6	140.7	8.8
Total Percent	689.4	43.1	910.6	56.9	1600	100

There is a strong relationship between apparent registrant type and Potentially Commercial activity, with a p-value for the relationship of less than .0001. Overall, 56.9 percent of domains show Potentially Commercial activity, but this percentage is higher for any apparent registrant type other than unknown, which only shows Potentially Commercial activity for 31.6 percent. The differences between the other three apparent registrant types are not large.

C. Potentially Commercial Activity Variables

Apparent Domain User Type

Table C.1: Summary of Potentially Commercial Activity Variables by Apparent Domain User Type

Commercial Activity Variable	Percent Yes					
	Natural Person	Legal Person	Domain Parked	No Online Content	Unknown Type	p-value
E-Commerce	3.5	15.0	1.0	0.0	6.9	<.0001
Membership (Online Content)	0.0	3.0	1.9	0.0	1.7	0.0056
Membership (Offline Content)	1.2	7.7	0.3	0.0	4.1	<.0001
Promotional Content (Offline)	14.5	42.4	1.0	0.5	14.7	<.0001
Promotional Content (Online)	6.9	10.4	3.8	0.5	4.2	<.0001
Host Promotional Content (Online)	1.2	4.0	33.6	0.0	1.7	<.0001
Third Party Banner Ads	5.9	12.3	2.6	0.3	9.4	<.0001
Host Banner Ads	1.2	5.8	49.7	0.6	0.0	<.0001
Pay-Per-Click Ads	12.9	22.8	79.7	5.3	25.0	<.0001
Host Pay-Per-Click Ads	1.2	2.0	13.5	0.1	0.6	<.0001
Potentially Commercial Activity	36.8	79.8	90.3	6.7	46.2	<.0001
Excluding Pay-Per-Click	31.0	72.1	61.9	1.8	30.0	<.0001

All of the p-values are less than 0.0001, indicating that there are very significant differences among the apparent domain user types on the potentially commercial activity variables. All potentially commercial activity variables are significantly more likely among legal persons, except for host banner ads and the two pay-per-clicks variables, where the highest potentially commercial activity is among the domain parked domains.

Table C.2: Potentially Commercial Activity by Apparent Domain User Type Weighted Cross-classified Frequency Counts

Potentially Commercial Activity	Apparent Domain User Type											
	Natural Person		Legal Person		Domain Parked		No Online Content		Unknown Type		Total	Percent
Not Detected	56.2	8.2	119.1	17.3	31.8	4.6	384.9	55.8	97.4	14.1	689.4	43.1
Detected	32.7	3.6	469.3	51.5	297.4	32.7	27.6	3.0	83.7	9.2	910.6	56.9
Total Percent	88.9	5.6	588.4	36.8	329.2	20.6	412.5	25.8	181.1	11.3	1600	100

Table C.2 is the transpose of Table A.2, showing how the apparent domain user distribution differs whether the domain shows Potentially Commercial activity or not. There is a strong relationship between apparent domain user type and Potentially Commercial activity, with a p-value for the relationship of less than .0001. Overall, 5.6 percent of the domain users are apparently natural persons, but this percentage is 3.6 percent for domains with Potentially Commercial activity and 8.2 for those without Potentially Commercial activity. Overall, 36.8 percent of the domain users are apparently legal persons, but this percentage is 51.5 percent for domains with Potentially Commercial activity and only 17.3 for those without Potentially Commercial activity. Overall, 20.6 percent of the domain users are parked domains, but this percentage is 32.7 percent for domains with Potentially Commercial activity and only 4.6 for those without Potentially Commercial activity. Overall, 25.8 percent of the domain users had no online content, but this percentage is only 3.0 percent for domains with Potentially Commercial activity and 55.8 for those without Potentially Commercial activity. Overall, 11.3 percent of the domain users were of an unknown type, but this percentage is 9.2 percent for domains with Potentially Commercial activity and only 14.1 for those without Potentially Commercial activity.

Apparent Registrant Type

Table C.3: Summary of Potentially Commercial Activity Variables by Apparent Registrant Type

Commercial Activity Variable	Percent Yes				
	Natural Person	Legal Person	Privacy/ Proxy	Unknown	p-value
E-Commerce	7.8	6.5	6.9	3.0	0.2383
Membership (Online Content)	0.9	1.9	1.6	3.7	0.1335
Membership (Offline Content)	2.8	3.5	5.2	1.5	0.1541
Promotional Content (Offline)	18.5	21.6	16.4	8.2	0.0019
Promotional Content (Online)	6.8	6.2	4.1	2.2	0.1044
Host Promotional Content (Online)	10.8	7.5	9.8	3.0	0.0173
Third Party Banner Ads	5.8	7.9	7.2	1.5	0.0365
Host Banner Ads	12.7	11.4	17.6	5.2	0.0017
Pay-Per-Click Ads	27.6	29.1	40.3	15.9	<.0001
Host Pay-Per-Click Ads	3.5	3.8	4.3	2.2	0.7255
Potentially Commercial Activity	54.6	59.5	64.0	31.6	<.0001
Excluding Pay-Per-Click	46.1	48.2	46.7	20.9	<.0001

Only five potentially commercial activity variables have p-values that indicate a significant different among the apparent registrants types (i.e., offline promo content, host online promotional content, third-party/host banner ads, pay-per-click ads). If the unknowns are ignored, there are three variables with statistically significant differences between registrants who are apparently natural or legal persons on the one hand and privacy/proxy registered domains on the other hand. The privacy/proxy registered domains have a statistically significantly less online promotional content, but statistically significantly more host banner ads and pay-per-click ads.

Table C.4: Potentially Commercial Activity by Apparent Registrant Type Weighted Cross-classified Frequency Counts

Potentially Commercial Activity	Apparent Registrant Type									
	Natural Person		Legal Person		Privacy/Proxy		Unknown		Total	Percent
Not Detected	229.6	33.3	249.5	36.2	114.0	16.5	96.2	14.0	689.4	43.1
Detected	285.6	31.3	372.3	40.9	208.3	22.9	44.4	4.9	910.6	56.9
Total Percent	515.3	32.2	621.8	38.9	322.3	20.1	140.7	8.8	1600	100

Table C.4 is the transpose of Table B.2, showing how the apparent registrant distribution differs whether the domain shows Potentially Commercial activity or not. There is a strong relationship between

apparent registrant type and Potentially Commercial activity, with a p-value for the relationship of less than .0001. Overall, 32.2 percent of the registrants are apparently natural persons, and this percentage differs little for domains with Potentially Commercial activity (31.3) and those without Potentially Commercial activity (33.3). Overall, 38.9 percent of the registrants are apparently legal persons, and this percentage differs little for domains with Potentially Commercial activity (40.9) and those without Potentially Commercial activity (36.2). Overall, 20.1 percent of the registrants are privacy/proxy registered domains, but this percentage is 22.9 percent for domains with Potentially Commercial activity and only 16.5 for those without Potentially Commercial activity; this difference is statistically significant. Overall, 8.8 percent of the registrants were of an unknown type, but this percentage is only 4.9 percent for domains with Potentially Commercial activity and 14.0 for those without Potentially Commercial activity.

D. Business Structure of Domain User

Generic business structure of the domain user was coded based on observed domain content that included HTML content and images extracted from “www.domainname”. Coders made direct observations on the domain user's business structure and indirect observations on other aspects, such as the domain user's business function, that may provide additional clues to the domain user's business structure. Their recorded observations were then categorized into 11 major types as described below. We searched in the coder observations for keywords that best characterize each category. When a record is associated with keywords corresponding to multiple business structure types, certain rules were applied to finalize it to a best fit category. Less than 3 percent of cases that were not suitable for automation were manually reviewed and finalized. Our main goal for this variable was to determine if the domain user could be considered a for-profit business, a non-profit business, or not a business at all. We split the for-profit businesses into sole proprietorships, partnerships, and corporations if we could. Some domains in languages other than English were clearly businesses, but were not classifiable. Domains with no content, as well as parked domains and under construction domains were assigned to unclear business structure categories. One other category was created for when some business activity was detected, but it was not clear whether or not the domain was a business. The remaining domains with no clear domain user type were assigned to an Undetermined category. Table D.1 shows the full frequency for the generic business structure of the domain user:

Table D.1: Generic Business Structure of Domain User

Description	Frequency	Percent
Undetermined	940	58.9
For Profit: Corporation	268	16.8
Not a Business (natural person, blog)	102	6.2
Unclear Business Structure: No Content (domain parked, under construction)	62	3.9
Unclear Business Structure: Unable to determine	49	3.1
For Profit: Partnership	38	2.4
Not For Profit (Nonprofit, governments, political, education, religious, or community groups)	37	2.3
For Profit: Sole Proprietor	32	2.0
Non-U.S. Business	31	1.9
For Profit: Other	29	1.8
Unclear Formal Structure: Apparent Business Activities	12	0.8

For analysis, we combined the eleven categories shown in Table D.1 into four categories. We combined all four "For Profit" categories together, kept the "Non-Profit" and "Not a Business" categories, and combined the other five categories into "Unclear Business Structure."

Therefore, our commercial business analysis variable has four levels as shown in Table D.2.

Table D.2: Business Structure of Domain User Variable Used in Analyses

Description	Frequency	Percent
Domain User appears to be for-Profit Business	410	25.6
Domain User appears to be non-Profit Business	37	2.3
Domain User is not a Business	102	6.4
Domain User has unclear Business Structure	1,051	65.7

Apparent Domain User Type

Table D.3: Apparent Domain User Type by Business Structure Weighted Cross-classified Frequency Counts

Apparent Domain User Type	Business Structure of Domain User									
	For Profit		Non-profit		Not a Business		Unclear Business Structure		Total	Percent
Natural Person	0	0	0	0	88.9	100	0	0	88.9	5.6
Legal Person	410.4	69.7	39.2	6.7	14.7	2.5	124.2	21.1	588.4	36.8
Domain Parked	0	0	0	0	0	0	329.2	100	329.2	20.6
No Online Content	0	0	0	0	0	0	412.5	100	412.5	25.8
Unknown Type	0	0	0	0	1.1	0.6	180.0	99.4	181.1	11.3
Total Percent	410.4	25.6	39.2	2.5	104.6	6.5	1045.9	65.4	1600	100

The relationship between apparent domain user type and the domain user's business structure is statistically significant with a chi-squared p-value of less than 0.0001. All of the For Profit and Non-Profit businesses have been classified as legal person domain users while all of the apparently natural person domain users have been classified as not a business. All of the domains parked and domains with no online content have an unclear business structure, while almost all of the unknown type domain users also have an unclear business structure. Looking at the row with domain users who are apparently legal persons, almost 70 percent of the domains appear to be for-profit businesses, while under 7 percent appear to be non-profit businesses and only 2.5 percent do not appear to be businesses at all. It should be noted that the sample size of the apparently non-profit business category is too small for analysis.

Apparent Registrant Type

Table D.4: Apparent Registrant Type by Business Structure Weighted Cross-classified Frequency Counts

Apparent Registrant Type	Business Structure of Domain User									
	For Profit		Non-profit		Not a Business		Unclear Business Structure		Total	Percent
Natural Person	100.4	19.5	12.7	2.5	60.9	11.8	341.2	66.2	515.3	32.2
Legal Person	229.7	36.9	20.1	3.2	15.0	2.4	357.0	57.4	621.8	38.9
Privacy/Proxy	62.4	19.4	6.4	2.0	23.4	7.3	230.1	71.4	322.3	20.1
Unknown	17.8	12.7	0	0	5.2	3.7	117.6	83.6	140.7	8.8
Total Percent	410.4	25.6	39.2	2.5	104.6	6.5	1045.9	65.4	1600	100

The relationship between apparent domain registrant type and business structure of domain user is statistically significant with a chi-squared p-value of less than 0.0001. Overall, 25.6 percent of the domain users have a for-profit business structure, but this percentage is 36.9 for domains registered by apparently legal persons. Except for unknown registrant types, two or three percent of the domain users have a non-profit business structure. Only 6.5 percent of the domains are used by an entity that could be classified as a non-business, but this percentage is almost double (11.8 percent) for domains registered to apparently natural persons and less than half (2.4 percent) for domains registered to apparently legal persons. Most of the domains in all registrant types, though, do have an unclear domain user's business structure.

Potentially Commercial Activity

Table D.5: Potentially Commercial Activity by Business Structure of Domain User

	Percent Yes				
	For Profit	**Non-Profit**	**Not a Business**	**Unclear Business Structure**	**p-value**
Potentially Commercial Activity	83.8	53.8	39.3	48.2	<.0001

The relationship Potentially Commercial activity and domain user's business structure is statistically significant with a chi-squared p-value of less than 0.0001. The For-Profit business structure domains had the highest percentage of Potentially Commercial activity (83.8 percent)[2], but the other business structures also showed a lot of Potentially Commercial activity (overall, 56.9 percent of the domains have shown Potentially Commercial activity).

[2] Note that business structure was coded independently of potentially commercial activity, so the presence of potentially commercial activity is not the reason a domain user was classified as a for-profit business.

E. Domain Name Extension (gTLD)

Table 1 above shows the top five generic top-level domains and the distribution of the domains in our sample across these gTLDs. We compare all five domain name extensions as much as possible below.

Apparent Domain User Type

Table E.1: Apparent Domain User Type by Domain Name Extension Weighted Cross-classified Frequency Counts

Apparent Domain User Type	Domain Name Extension											
	*.com		*.net		*.org		*.info		*.biz		Total	Percent
Natural Person	68.5	77.0	13.5	15.2	2.2	2.4	3.9	4.4	0.8	0.9	88.9	5.6
Legal Person	451.9	76.8	58.3	9.9	46.5	7.9	22.6	3.8	9.1	1.5	588.4	36.8
Domain Parked	246.5	74.9	36.4	11.1	21.6	6.6	19.7	6.0	4.9	1.5	329.2	20.6
No Online Content	281.3	68.2	50.0	12.1	32.4	7.9	41.3	10.0	7.5	1.8	412.5	25.8
Unknown Type	140.1	77.4	13.5	7.5	13.0	7.2	10.8	6.0	3.6	2.0	181.1	11.3
Total Percent	1188.2	74.3	171.8	10.7	115.7	7.2	98.3	6.1	26.0	1.6	1600	100

There is enough of a relationship between apparent domain user type and generic top-level domain (gTLD) name extension for a significant chi-square p-value of 0.0381. However, it does not appear to be a strong relationship. Overall, 74.3 percent of all domains are *.com domains, and only the No Online Content domains differ (68.2 percent). Overall, 10.7 percent of all domains are *.net, with the highest rate among the domain users who are apparently natural persons (15.2 percent) and the lowest rate among the unknown domain user types (7.5 percent). About seven percent of all domain user types are *.org except the apparently natural person domain users (2.4 percent). The most variable rates occur for the *.info gTLD. Overall, 6.1 percent of the domains are *.info domains, but the no online content domains have a 10.0 percent rate while the apparently natural person domain users (4.4 percent) and apparently legal person domain users (3.8 percent) have lower rates. The *.biz gTLD represents about one or two percent of domains in all domain user types.

Apparent Registrant Type

Table E.2: Apparent Registrant Type by Domain Name Extension
Weighted Cross-classified Frequency Counts

Apparent Registrant Type	Domain Name Extension											
	*.com		*.net		*.org		*.info		*.biz		Total	Percent
Natural Person	381.3	74.0	55.2	10.7	34.6	6.7	32.4	6.3	11.7	2.3	515.3	32.2
Legal Person	455.1	73.2	76.0	12.2	57.3	9.2	24.6	4.0	8.8	1.4	621.8	38.9
Privacy/Proxy	238.1	73.9	29.2	9.0	19.5	6.0	31.5	9.8	4.2	1.3	322.3	20.1
Unknown	113.8	80.9	11.5	8.1	4.3	3.1	9.8	7.0	1.3	0.9	140.7	8.8
Total Percent	1188.2	74.3	171.8	10.7	115.7	7.2	98.3	6.1	26.0	1.6	1600	100

The relationship between domain name extensions and apparent registrant type is significant with a chi-squared p-value of 0.0124. Overall, 74.3 percent of all domains are *.com domains, and only the Unknown Registrant Type domains differ (80.9 percent). Overall, 10.7 percent of all domains are *.net, with the highest rate among the registrants who are apparently legal persons (12.2 percent) and the lowest rates among the privacy/proxy registered domains (9.0 percent) and the unknown registrant types (8.1 percent). Overall, 7.2 percent of all domains are *.org domains, but the percentage of registrants that are apparently legal persons is 9.2 while the percentage for Unknown Registrant Types is only 3.1 percent. Overall, 6.1 percent of the domains are *.info domains, but the privacy/proxy registered domains have a 9.8 percent rate while the apparently legal person registrants only have a 4.0 percent rate. Overall, 1.6 percent of all domains are *.biz domains, but this rate is higher for registrants who are apparently natural persons (2.3 percent) and lower for Unknown Registrant Types (0.9 percent).

Potentially Commercial Activity

Table E.3: Potentially Commercial Activity by Domain Name Extension

	Percent Yes					
	*.com	*.net	*.org	*.info	*.biz	p-value
Potentially Commercial Activity	59.0	55.8	47.7	47.0	50.0	.0315

The relationship Potentially Commercial activity and domain name extension is statistically significant with a chi-squared p-value of 0.0315. Compared with other tables, the differences are not that large, but the *.com and *.net domains do show more Potentially Commercial Activity than the *.org and *.info domains.

F. Registrant Country/Region of the World

Through our research, we were able to identify the registrant country for all but 82 of the domain names. For one domain name, there was conflicting information as to whether it was in Japan or Australia; for the remaining 81 missing registrant countries, no WHOIS information existed to be used to determine the registrant country. Table F.1 shows the countries represented by at least one domain name in our sample.

Table F.1: Countries Represented in the Registrant ID Study Domain Sample

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
United States	864	54.0	864	54.0
China	76	4.8	940	58.8
United Kingdom	76	4.8	1,016	63.5
Germany	56	3.5	1,072	67.0
Australia	50	3.1	1,122	70.1
Canada	50	3.1	1,172	73.3
Spain	34	2.1	1,206	75.4
France	31	1.9	1,237	77.3
Japan	29	1.8	1,266	79.1
The Netherlands	26	1.6	1,292	80.8
Italy	22	1.4	1,314	82.1
Turkey	20	1.3	1,334	83.4
India	17	1.1	1,351	84.4
Switzerland	11	0.7	1,362	85.1
Russia	11	0.7	1,373	85.8
Indonesia	9	0.6	1,382	86.4
Brazil	8	0.5	1,390	86.9
Hong Kong	8	0.5	1,398	87.4
Vietnam	8	0.5	1,406	87.9
Singapore	7	0.4	1,413	88.3
Belgium	6	0.4	1,419	88.7
Cayman Islands	6	0.4	1,425	89.1
Norway	6	0.4	1,431	89.4
Sweden	6	0.4	1,437	89.8
Thailand	6	0.4	1,443	90.2
Czech Republic	4	0.3	1,447	90.4
Ireland	4	0.3	1,451	90.7
South Korea	4	0.3	1,455	90.9
Mexico	4	0.3	1,459	91.2
South Africa	4	0.3	1,463	91.4
Bermuda	3	0.2	1,466	91.6

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Denmark	3	0.2	1,469	91.8
Finland	3	0.2	1,472	92.0
Greece	3	0.2	1,475	92.2
Philippines	3	0.2	1,478	92.4
Poland	3	0.2	1,481	92.6
Saudi Arabia	3	0.2	1,484	92.8
Bosnia and Herzegovina	2	0.1	1,486	92.9
Hungary	2	0.1	1,488	93.0
Israel	2	0.1	1,490	93.1
Iran	2	0.1	1,492	93.3
Malaysia	2	0.1	1,494	93.4
New Zealand	2	0.1	1,496	93.5
Venezuela	2	0.1	1,498	93.6
British Virgin Islands	2	0.1	1,500	93.8
United Arab Emirates	1	0.1	1,501	93.8
Argentina	1	0.1	1,502	93.9
Austria	1	0.1	1,503	93.9
Bolivia	1	0.1	1,504	94.0
Bahamas	1	0.1	1,505	94.1
Chile	1	0.1	1,506	94.1
Cyprus	1	0.1	1,507	94.2
Egypt	1	0.1	1,508	94.3
Croatia	1	0.1	1,509	94.3
Jordan	1	0.1	1,510	94.4
Lebanon	1	0.1	1,511	94.4
Nicaragua	1	0.1	1,512	94.5
Peru	1	0.1	1,513	94.6
Puerto Rico	1	0.1	1,514	94.6
Qatar	1	0.1	1,515	94.7
Serbia	1	0.1	1,516	94.8
Ukraine	1	0.1	1,517	94.8
Uruguay	1	0.1	1,518	94.9
Ambiguous	1	0.1	1,519	94.9
Unknown (no data available)	81	5.1	1,600	100.0

For countries with at least fifty (50) domain names (United States, China, United Kingdom, Germany, Australia, and Canada), we have analyzed them separately. We have combined the other countries by region as follows: Other Europe, Other Asia/Pacific, and Other (North America excluding the U.S. and Canada, South America, Caribbean Islands, and Africa). Table F.2 shows the frequency for the analysis variable we used to represent country/region of the world. We concentrated on the nine subgroups with data available to analyze.

Table F.2: Countries/Regions of the World Used in Analyses

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
United States	864	54.0	864	54.0
China	76	4.8	940	58.8
United Kingdom	76	4.8	1,016	63.5
Germany	56	3.5	1,072	67.0
Australia/New Zealand	52	3.3	1,124	70.3
Canada	50	3.1	1,174	73.4
Other Europe	170	10.6	1,344	84.0
Other Asia/Pacific	136	8.5	1,480	92.5
Other	38	2.4	1,518	94.9
Ambiguous/Missing	82	5.1	1,600	100.0

Apparent Domain User Type

Table F.3: Apparent Domain User Type by Country/Region of the World Weighted Cross-classified Frequency Counts

Apparent Domain User Type	Registrant Country											
	United States		China		United Kingdom		Germany		Australia/ New Zealand		Canada	
Natural Person	44.4	50.0	2.1	2.4	7.3	8.2	8.6	9.7	2.1	2.4	0	0
Legal Person	320.7	54.9	33.7	5.8	25.5	4.4	24.8	4.2	16.7	2.9	19.2	3.3
Domain Parked	217.9	67.3	11.6	3.6	14.9	4.6	6.3	1.9	16.1	5.0	13.7	4.2
No Online Content	202.0	59.3	20.2	5.9	11.1	3.3	9.3	2.7	13.1	3.8	9.8	2.9
Unknown Type	85.2	47.9	11.6	6.5	16.0	9.0	5.5	3.1	4.1	2.3	8.4	4.7
Total Percent	870.3	57.4	79.2	5.2	74.9	4.9	54.6	3.6	52.2	3.4	51.1	3.4

Apparent Domain User Type	Other Europe		Other Asia		Other		Total	Percent
Natural Person	11.8	13.3	12.5	14.1	0	0	88.9	5.9
Legal Person	75.3	12.9	49.6	8.5	18.7	3.2	584.2	38.5
Domain Parked	19.1	5.9	18.0	5.6	6.3	2.0	323.9	21.4
No Online Content	37.1	10.9	29.8	8.7	8.5	2.4	340.8	22.5
Unknown Type	18.4	10.3	24.3	13.7	4.3	2.4	177.9	11.7
Total Percent	161.6	10.7	134.2	8.9	37.8	2.4	1516	100

There is a strong relationship between apparent domain user type and country/region of the world, with a p-value of less than .0001. Overall, 57.4 percent of the domains have a United States registrant, but this percentage is 67.3 percent for parked domains and is only 50.0 percent for domain users that are

apparently natural persons (and 47.9 percent for unknown domain user types). Overall, 5.2 percent of the domains have Chinese registrants, but this percentage is 6.5 percent for unknown domain user type domains while this percentage is only 2.4 percent for domain users that are apparently natural persons (and 3.6 percent for parked domains). Overall, 4.9 percent of the domains have United Kingdom registrants, but this percentage is 8.2 for domain users that are apparently natural persons (and 9.0 percent for unknown domain user types) while this percentage is only 3.3 percent for domains with no online content. Overall, 3.6 of the domains have a German registrant, but this percentage is 9.7 percent for domain users that are apparently natural persons and is only 1.9 percent for parked domains. Overall, 3.4 of the domains have an Australia or New Zealand registrant, but this percentage is 5.0 percent for parked domains and is only 2.4 percent for domain users that are apparently natural persons (and 2.3 for unknown domain user types). Canadian registrants make up three to five percent of the domains in each domain user type category except that there are no Canadian registrants for domain users that are apparently natural persons. The overall percentage for other European countries (besides the United Kingdom and Germany) is 10.7, but this percentage is higher for domain users that are apparently natural persons (13.3 percent) and domain users that are apparently legal persons (12.9 percent), but lower for parked domains (5.9 percent). The overall percentage for other Asian and Pacific countries (besides China) is 8.9, but this percentage is 14.1 percent for domain users that are apparently natural persons (and 13.7 percent for unknown domain user types), but lower for parked domains (5.6 percent). Registrants from other countries and regions make up two to three percent of the domains in each domain user type category except that there are no registrants from these other countries/regions for domain users that are apparently natural persons.

It seems clear from the above that domain users who are apparently natural persons differ the most from the other categories in the distribution by country/region of the world.

Apparent Registrant Type

Table F.4: Apparent Registrant Type by Country/Region of the World Weighted Cross-classified Frequency Counts

Apparent Registrant Type	Registrant Country											
	United States		China		United Kingdom		Germany		Australia/ New Zealand		Canada	
Natural Person	239.9	46.8	48.7	9.5	33.3	6.5	27.8	5.4	8.7	1.7	11.6	2.3
Legal Person	366.7	59.4	15.8	2.6	33.0	5.3	25.7	4.2	12.5	2.0	19.2	3.1
Privacy/Proxy	238.0	74.3	3.1	1.0	3.4	1.1	1.0	0.3	29.9	9.3	20.3	6.3
Unknown	25.7	39.0	11.6	17.6	5.3	8.0	0	0	1.1	1.6	0	0
Total Percent	870.3	57.4	79.2	5.2	74.9	4.9	54.6	3.6	52.2	3.4	51.1	3.4

Apparent Registrant Type	Other Europe		Other Asia		Other		Total	Percent
Natural Person	72.6	14.2	58.0	11.3	11.5	2.1	512.1	33.8
Legal Person	72.9	11.8	54.1	8.8	17.6	2.9	617.6	40.7
Privacy/Proxy	5.5	1.7	14.8	4.6	4.2	1.3	320.2	21.1
Unknown	10.7	16.2	7.4	11.2	4.3	6.4	65.8	4.3
Total Percent	161.6	10.7	134.2	8.9	37.6	2.4	1516	100

There is a strong relationship between apparent registrant type and country/region of the world, with a p-value of less than .0001. Overall, 57.4 of the domains have a United States registrant, but this percentage is 74.3 percent for privacy/proxy registered domains and is only 46.8 percent for registrants that are apparently natural persons (and 39.0 percent for unknown registrant types). Overall, 5.2 percent of the domains have Chinese registrants, but this percentage is 17.6 percent for unknown registrant type domains while this percentage is only 1.0 percent for privacy/proxy registered domains (and only 2.6 percent for registrants that are apparently legal persons). Overall, 4.9 percent of the domains have United Kingdom registrants, but this percentage is 6.5 for registrants that are apparently natural persons (and 8.0 percent for unknown registrant types) while this percentage is only 1.1 percent for privacy/proxy registered domains. Overall, 3.6 of the domains have a German registrant, but almost all are registrants that are apparently natural persons (5.4 of apparently natural person registrants) and registrants that are apparently legal persons (4.2 percent of all apparently legal person registrants) while almost none are privacy/proxy registered domains (0.3 percent of privacy/proxy registered domains) or unknown registrant types (none of the 66 unknown registrant type registrants). Overall, 3.4 of the domains have an Australia or New Zealand registrant, but this percentage is 9.3 percent for privacy/proxy registered domains and two percent for all other apparent registrant types). Overall, 3.4 of the domains have a

Canadian registrant, but this percentage is 6.3 percent for privacy/proxy registered domains while there were no unknown registrant cases (out of 66 total unknown registrant cases) with Canadian registrants. The overall percentage for other European countries (besides the United Kingdom and Germany) is 10.7, but this percentage is much lower for privacy/proxy registered domains (1.7 percent) and higher for registrants that are apparently natural persons (14.2 percent) and unknown registrant type registrants (16.2 percent). The overall percentage for other Asian and Pacific countries (besides China) is 8.9, but this percentage is 11.3 percent for registrants that are apparently natural persons (and 11.2 percent for unknown registrant types), but lower for privacy/proxy registered domains (4.6 percent). The overall percentage for all other countries and regions is 2.4, but this percentage is higher (6.4 percent) for unknown registrant types and lower (1.3 percent) for privacy/proxy registered domains.

It seems clear from the above that privacy/proxy registered domains differ the most from the other categories in the distribution by country/region of the world.

Potentially Commercial Activity

Table F.5: Potentially Commercial Activity by Country/Region

	Percent Yes					
	United States	**China**	**United Kingdom**	**Germany**	**Australia/ New Zealand**	**Canada**
Potentially Commercial Activity	63.8	50.5	62.6	39.1	58.9	60.3

	Percent Yes			
	Other Europe	**Other Asia**	**Other**	**p-value**
Potentially Commercial Activity	51.7	50.3	69.1	.0003

There is a strong relationship between apparent registrant type and country/region of the world, with a p-value of .0003. Ignoring the "Other" category, the United States has the highest rate of Potentially Commercial activity (63.8 percent) while Germany has the lowest rate (39.1 percent). The United Kingdom has the second highest rate (62.6 percent) while China and the Other Asia region have rates around 50 percent.

G. Relationship of Domain User to Registrant

The relationship between the Domain User and the Registrant was coded during the second phase of the Domain User manual coding process. The entity listed in the WHOIS data Registrant Name and Registrant Organization fields were compared to the Domain User and the type of the relationship existing between the two entities was recorded. Here is a frequency:

Table G.1: Relationship between Domain User and Registrant

Relationship Description	Frequency	Percent
No Apparent Relationship: Unable to determine relationship	868	54.3
Domain User is Customer of Registrant: Privacy or Proxy service registered domain	327	20.4
Domain User same as Registrant both Legal Person	198	12.4
Domain User is Employer of Registrant	79	4.9
Domain User same as Registrant, both Natural Person	67	4.2
Domain User is Customer of Registrant: Web Developer/Development /Consulting company registered domain	27	1.7
Domain User is Customer of Registrant: Hosting or Domain provider	19	1.2
Other Specify	13	0.8
No Apparent Relationship: Registrant appears fictitious or falsified	2	0.1

For our analysis purposes, we collapsed these nine categories into four categories with the Other Specify categorized based on the text description. We combined the two categories where the Domain User is also the Registrant, whether Natural or Legal person (plus three Other Specify cases); we kept the Domain User is Customer of Privacy/Proxy Registered Domain separate, but we combined the two other "Domain User is Customer" categories together (plus one Other Specify case); and we combined the "Domain User is Employer" category with two Other Specify cases where the Domain User was the Employee of the Registrant. The remaining two "No Apparent Relationship" categories were combined with the remaining seven Other Specify cases to make the "Unknown" category. Table G.2 shows the frequency of the Relationship variable used in our analyses:

Table G.2: Relationship Variable Used in Analyses

Relationship Description	Frequency	Percent
Domain User Same as Registrant	268	16.8
Domain User is Customer of Privacy/Proxy Registered Domain (PRIVACY/PROXY)	327	20.4
Domain User is Customer of Other Registrant (OTHER CUSTOMER)	47	3.0
Domain User is Employer/Employee of Registrant (EMPLOYER/EMPLOYEE)	81	5.1
Unable to Determine Relationship	877	54.8

Apparent Domain User Type

Table G.3: Relationship of Domain User and Registrant by Domain User Type Weighted Cross-classified Frequency Counts

	Relationship of Domain User to Registrant											
Apparent Domain User Type	**Domain User Same as Registrant**		**Privacy/ Proxy**		**Other Customer**		**Employer/ Employee**		**Unable to Determine Relationship**		**Total**	**Percent**
Natural Person	62.0	69.8	19.2	21.6	2.1	2.4	1.1	1.2	4.5	5.0	88.9	5.6
Legal Person	208.6	35.5	76.5	13.0	37.6	6.4	75.7	12.9	190.0	32.3	588.4	36.8
Domain Parked	3.2	1.0	122.2	37.1	3.2	1.0	0	0	200.6	61.0	329.2	20.6
No Online Content	0	0	71.9	17.4	1.1	0.3	0	0	339.5	82.3	412.5	25.8
Unknown Type	0	0	38.1	21.1	4.5	2.5	2.1	1.2	136.4	75.3	181.1	11.3
Total Percent	273.8	17.1	327.9	20.5	48.4	3.0	78.9	4.9	871.0	54.4	1600	100

The relationship between apparent domain user type and the relationship of domain user to registrant is highly significant with a chi-squared p-value of less than 0.0001. Overall, the percentage of domain users who are the same entity as the registrant is 17.1, but this percentage is much higher for domain users who are apparently natural persons (69.8 percent) and domain users who are apparently legal persons (35.5 percent) while very few for the less defined domain user types (parked domains, no online content and unknown domain user type). Overall, the percentage of domain users who are clients of privacy/proxy registered domains is 20.5 percent, but this percentage is higher for parked domains (37.1 percent) and lower for domains with no online content (17.4 percent) and domain users who are apparently legal persons (13.0 percent). Overall, the percentage of domain users who are clients of other registrants (not privacy/proxy

registered domains) is 3.0 percent, but this percentage is higher for domain users who are apparently legal persons (6.4 percent) and lower for domains with no online content (0.3 percent) and parked domains (1.0 percent). Almost all of the employer/employee relationships between the domain user and registrant were for domain users who are apparently legal persons (12.9 percent of domain users who are apparently legal persons), with all other domain user types having such a relationship only zero or one percent of the time. Overall, we were unable to determine the relationship for 54.4 of the domains, but this percentage was especially low (5.0 percent) for domain users who are apparently natural persons, lower (32.3 percent) for domain users who are apparently legal persons and highest for domains with no online content (82.3 percent) and unknown domain user type domains (75.3 percent).

Apparent Registrant Type

Table G.4: Relationship of Domain User and Registrant by Registrant Type Weighted Cross-classified Frequency Counts

Apparent Registrant Type	Relationship of Domain User to Registrant											
	Domain User Same as Registrant		Privacy/ Proxy		Other Customer		Employer/ Employee		Unable to Determine Relationship		Total	Percent
Natural Person	88.1	17.1	9.6	1.9	7.3	1.4	35.8	6.9	374.5	72.7	515.3	32.2
Legal Person	173.1	27.8	16.0	2.6	31.7	5.1	41.0	6.6	360.0	57.9	621.8	38.9
Privacy/Proxy	3.2	1.0	299.2	92.8	8.3	2.6	2.1	0.7	9.5	2.9	322.3	20.1
Unknown	9.5	6.7	3.1	2.2	1.0	0.7	0	0	127.0	90.3	140.7	8.8
Total Percent	273.8	17.1	327.9	20.5	48.4	3.0	78.9	4.9	871.0	54.4	1600	100

The relationship between apparent registrant type and relationship of domain user to registrant is highly significant with a chi-squared p-value of less than 0.0001. We expect that the privacy/proxy registered domains will have their domain users all be customers, and this is almost true. Overall, the percentage of domain users who are the same entity as the registrant is 17.1, but this percentage is 27.8 for domain users who are apparently legal persons, while this percentage is only 1.0 percent for privacy/proxy registered domains (and is only 6.7 percent for unknown registrant type registrants). Overall, the percentage of domain users who are customers of privacy/proxy registered domains is 20.5 percent, but this percentage is much higher for privacy/proxy registered domains (92.8 percent) and much lower (less than eight percent) for all three of the other registrant types. Overall, the percentage of domain users who are customers,

but are not privacy/proxy registered domains, is 3.0 percent, but this percentage is much higher for domain users who are apparently legal persons (5.1 percent) and lower for domain users who are apparently natural persons (1.4 percent). Overall, the percentage of domains with an employer/employee relationship between the domain user and registrant was 4.9 percent, but almost all of these relationships were for domain users who are apparently natural persons (6.9 percent of domain users who are apparently natural persons) and for domain users who are apparently legal persons (6.6 percent of domain users who are apparently legal persons) with privacy/proxy registered domains and unknown registrant types having such a relationship less than one percent of the time. Overall, we were unable to determine the relationship for 54.4 of the domains, but this percentage was especially low (2.4 percent) for privacy/proxy registered domains and higher for registrants who are apparently natural persons (72.7 percent) and for unknown registrant types (90.3 percent).

Potentially Commercial Activity

Table G.5: Potentially Commercial Activity by Relationship between Registrant and the Domain User

	Percent Yes					
	Domain User Same as Registrant	**Privacy/ Proxy**	**Other Customer**	**Employer/ Employee**	**Unable to Determine Relationship**	**p-value**
Potentially Commercial Activity	67.5	65.9	80.4	83.5	46.5	<.0001

The relationship between Potentially Commercial activity and the relationship between the registrant and the domain user is statistically significant with a chi-squared p-value of less than 0.0001. The relationships that showed the most Potentially Commercial activity occurs when the user and registrant have an employer/employee relationship or a (non-privacy/proxy) customer relationship, while the lowest Potentially Commercial activity rate was among those domains where we were unable to determine the relationship between the domain user and registrant. This low rate may be related to the fact that we weren't able to determine the relationship for domains with no online content.

H. Other Coded Behavior Variables

Two other coded behavior variables were used to indicate whether any alleged illegal or harmful activity was detected and whether any explicit sexual imagery was found (this differs from the analysis below on whether a domain could be matched to any blacklists). These allegedly illegal or harmful activities were coded during the Domain Content manual coding process by manually reviewing the web content for evidence of each of the activities listed in Table H.1. During the training process, coders were supplied with definitions of each of the activities, and a few examples of websites engaging in the activities were provided. However, it should be noted that the coders were not experts in Internet crime and detecting the presence of these activities on web pages. Table H.1 shows the frequency of our allegedly illegal or harmful activity variable:

Table H.1: Allegedly Illegal or Harmful Activities: Manually Coded

Allegedly illegal or harmful Activity	Frequency	Percent
No allegedly illegal or harmful activities detected	1,582	98.9
Spam	4	0.3
Advance fee fraud (aka 419 scams)	4	0.3
Phishing	3	0.2
Cybersquatting/Typosquatting	3	0.2
Counterfeit merchandise (i.e., domain website appears to sell CM)	2	0.1
Trademark infringement (i.e., domain website appears to…)	1	0.1
Malware	1	0.1
Intellectual property theft	0	0.0
Child sexual images	0	0.0
Identity theft	0	0.0
Money laundering	0	0.0

Allegedly illegal or harmful activities were only observed for 18 out of the 1,600 domains (1.1 percent). In our analyses, we converted this variable to a binary variable of whether any alleged illegal activity was detected. Table H.2 shows the frequency of whether explicit sexual images were at the domain:

Table H.2: Explicit Sexual Images: Manually Coded

Explicit Sexual Images	Frequency	Percent
No	1,584	99.0
Yes	16	1.0

Even though both of these variables were rarely yes, we still carried out analyses to see if these two behaviors were more likely among certain subgroups.

Apparent Domain User Type

Table H.3: Coded Behavior Variables by Apparent Domain User Type

Coded Variable	Percent Yes					p-value
	Natural Person	Legal Person	Domain Parked	No Online Content	Unknown Type	
Allegedly illegal or harmful Activity	1.2	2.1	1.0	0	1.2	0.0653
Explicit Sexual Images	2.4	1.6	0.6	0	1.7	0.0611

While the p-values are close to significant, the p-values are not significant even though one of the apparent domain user types (no online content) could not show these coded behaviors. For allegedly illegal or harmful activity, there is a slightly higher rate (2.1 percent) among the domain users who are apparently legal persons. Few of the parked domains showed explicit sexual images (0.6 percent) while there was a slightly higher rate for domain users who are apparently natural persons.

Apparent Registrant Type

Table H.4: Coded Behavior Variables by Apparent Registrant Type

Coded Variable	Percent Yes				p-value
	Natural Person	Legal Person	Privacy/ Proxy	Unknown	
Allegedly illegal or harmful Activity	1.6	0.5	1.6	1.5	0.0580
Explicit Sexual Images	0.6	1.0	1.6	1.5	0.5173

The p-value for explicit sexual images shows no significant differences between the apparent registrant types. The p-value for allegedly illegal or harmful activity shows that the difference between registrants who apparently are legal persons (0.5 percent) and all other apparent

registrant types (1.5-1.6 percent) is almost statistically significant. Meanwhile, the percentages of explicit sexual images are lower for registrants who apparently are natural persons (0.6 percent) and for registrants who apparently are legal persons (1.0 percent), but the differences in the percentage of domains with explicit sexual images could be due to random error.

Potentially Commercial Activity

Table H.5: Coded Behavior Variables by Potentially Commercial Activity

Coded Variable	Percent Yes		p-value
	No Potentially Commercial Activity	Potentially Commercial Activity	
Allegedly illegal or harmful Activity	0.8	1.5	0.5509
Explicit Sexual Images	0.9	1.2	0.6416

For both of these coded behavior variables, the domains with Potentially Commercial activity have a higher rate of the coded behavior, but the differences are not large enough to be statistically significant.

I. Blacklist Variables

In an effort to determine allegedly illegal or harmful activities, DNSBL lists were scanned for each sample member. The DNSBL strategy was to obtain all the "A-RECORDS" associated with the domain for each sample member. For each A-RECORD, the returned IP address was checked against a series of DNSBLs. After running this process, we reviewed the frequency of responses received from each DNSBL. Many of the DNSBLs did not return a response, so they were removed from our analysis. For the remaining DNSBLs which returned a response, NORC conducted a review of the site to determine the relevancy of the list. Many of the lists contained an abundance of historic DNSBL listings or were no longer actively maintained, so these were removed from the analysis. Some of the response octates returned by the DNSBLs provided a trustworthiness score of the listing to indicate how sure the DNSBL is that the listing is accurate. Scores of low trustworthiness were removed from the analysis. Table I.1 is a summary of the allegedly illegal or harmful activity categories as determined by the top-ranked blacklists. It is possible for a domain to be categorized in more than one way, so the categories in the summary table are not distinct. The total number of domains associated with any top-ranked blacklist activity is provided at the bottom of the table.

Table I.1: Allegedly illegal or harmful Activities: Domains Found on Top-Ranked Blacklists

Description	Frequency	Percent*
Abusive	2	0.1
Abusive host	5	0.3
Abusive host & anonymous-state	28	1.8
Backscatter	28	1.8
Ddos attacks	1	0.1
Dynamic-ip	7	0.4
Spam	82	5.1
Spam abuse vulnerability	6	0.4
Spam bad host, no cookie	1	0.1
Suspicious	5	0.3
Suspicious & comment spammer	1	0.1
Tor network	1	0.1
Trojan/virus/bot	2	0.1
On Any Top-Ranked Blacklist	141	8.8

In the following analyses, we restrict our analyses to the most common four allegedly illegal or harmful activities: any of the top-ranked blacklists (141 cases), abusive host and anonymous-state (28 cases), backscatter (28 cases), and spam (82 cases).

Apparent Domain User Type

Table I.2: Summary of Blacklist Variables by Apparent Domain User Type

Blacklist Variable	Percent Yes					
	Natural Person	Legal Person	Domain Parked	No Online Content	Unknown Type	p-value
On Any Top-Ranked Blacklist	11.8	12.4	5.4	6.1	9.8	0.0009
Abusive host/anonymous	2.4	1.6	3.8	1.0	0.5	0.0290
Backscatter	3.5	2.9	0.3	1.0	1.2	0.0172
Spam	5.9	8.2	1.3	3.3	6.9	<.0001

All four blacklist variables show statistically significant differences between the apparent domain user types. Overall, 8.8 percent of domains appear on any top-ranked blacklist, but this percentage is higher for domains that are apparently legal persons (12.4 percent) and domains that are apparently natural

persons (11.8 percent). Parked domains (5.4 percent) and domains with no online content (6.1 percent) have the lowest rates of appearing on any top-ranked blacklist. For abusive host/anonymous blacklists, the parked domains have the highest rate (3.8 percent) of appearing on a blacklist of this type while domains with no online content (1.0 percent) and unknown domain user types (0.5 percent) have the lowest rates. For backscatter blacklists, the highest rates belong to domains that are apparently used by natural persons (3.5 percent) and domains that are apparently used by legal persons (2.9 percent), while the rates are around one percent or lower for the other three domain user types. For spam blacklists, the highest rate is for domains that are apparently used by legal persons (8.2 percent) while unknown domain user types (6.9 percent) and domains that are apparently used by natural persons (5.9 percent) also have higher rates than domains with no online content (3.3 percent) and parked domains (1.3 percent). Comparing just domains used by apparently natural persons with those that are used by apparently legal persons, they have similar overall rates of appearing on any top-ranked blacklist, but domains used by apparently legal persons have a higher spam blacklist rate while domains that are apparently used by natural persons have slightly higher rates in the two larger categories with enough positive matches to separate out (abusive host/anonymous and backscatter).

Apparent Registrant Type

Table I.3: Summary of Blacklist Variables by Apparent Registrant Type

Blacklist Variable	Percent Yes				
	Natural Person	Legal Person	Privacy/ Proxy	Unknown	p-value
On Any Top-Ranked Blacklist	11.6	8.0	7.9	6.7	0.0981
Abusive host/anonymous	2.2	2.0	1.3	0.7	0.5826
Backscatter	1.9	1.9	1.0	2.2	0.6971
Spam	7.9	3.9	3.9	4.4	0.0138

Overall, domains that are apparently registered by natural persons have a higher rate of appearing on any top-ranked blacklist than other registrant types, but the difference is not statistically significant. Domains that are apparently registered by natural persons do have a significantly higher rate of appearing on spam blacklists, however, with a rate (7.9 percent) that is about double the other registrant types (around four percent). The differences in abusive host/anonymous and backscatter blacklists are not significant, but the privacy/proxy registered domains have low rates for both.

Potentially Commercial Activity

Table I.4: Summary of Blacklist Variables by Potentially Commercial Activity

Blacklist Variable	Percent Yes		
	No Potentially Commercial Activity	Potentially Commercial Activity	p-value
On Any Top-Ranked Blacklist	8.3	9.5	0.3832
Abusive host/anonymous	1.4	2.2	0.2265
Backscatter	1.6	1.9	0.6433
Spam	4.6	5.7	0.3234

There are no significant differences in blacklist appearance between the domains with and without Potentially Commercial activity, but the rates are higher for domains with Potentially Commercial activity for all four variables shown.

J. Whitelist Variables

Similar to the blacklists consulted, we also checked all IPs associated with the A-RECORDS for the 1,600 domains against the whitelist hosted by www.dnswl.org and two additional whitelists. If a response was returned, this signified presence on a whitelist. The response octate of the dnswl.org gave additional information on the category of the entry on the whitelist. Table J.1 is a summary of the octate results returned by the whitelists. It is possible for a domain to be identified by more than one whitelist, so the categories in the summary table are not distinct. The total number of domains associated with any of the four whitelists is provided at the bottom of the table.

Table J.2: Domains Found on Whitelists

Description	Frequency	Percent*
Retail/Wholesale Serices	1	0.1
Service/Network Providers	130	8.1
Email Service Providers	2	0.1
No Whitelist Octate	96	6.0
On Any Whitelist	204	12.8

It is natural to wonder if any of the domains were found on any of the whitelists and any of the blacklists, so Table J.2 answers this question:

Table J.2: Domains Found on Whitelists and Blacklists
Weighted Cross-classified Frequency Counts

	On Any Whitelist					
On Any Blacklist	**No**		**Yes**		**Total**	**Percent**
No	1265.9		190.0		1455.9	91.0
Yes	130.5		13.6		144.1	9.0
Total Percent	1396.4	87.3	203.7	12.7	1600.0	100.0

According to Table J.2, almost one percent of the 1,600 domains were found on at least one top-ranked blacklist as well as at least one whitelist. Of the 204 domains matched to a whitelist, 6.7 percent also matched to a top-ranked blacklist compared to 9.3 percent of those that didn't match to a whitelist. Of the 141 domains matched to a top-ranked blacklist, 9.4 percent also matched to a whitelist compared to 13.1 percent of those that didn't match to a top-ranked blacklist.

Apparent Domain User Type

Table J.3: Domains Found on Whitelists by Apparent Domain User Type

Whitelist Variable	Percent Yes					
	Natural Person	Legal Person	Domain Parked	No Online Content	Unknown Type	p-value
On Any Whitelist	9.5	14.2	24.7	4.5	6.5	<.0001
Service/Network Providers	5.9	8.4	15.8	3.7	3.6	<.0001
No Whitelist Octate	3.6	8.0	11.4	1.1	2.9	<.0001

All three of these variables show highly significant differences. Parked domains have the highest rate of being on any whitelist, and they also have the highest rates in the two larger categories with enough positive matches to separate out (service/network providers and no whitelist octate). The next two highest rates for each of the three variables are for domains that are apparently used by legal persons and domains that are apparently used by natural persons. For all three variables, domains that are apparently used by legal persons have higher rates than domains that are apparently used by natural persons. The lowest rates for all three variables belong to domains with no online content and unknown domain user types.

Apparent Registrant Type

Table J.4: Domains Found on Whitelists by Apparent Registrant Type

Whitelist Variable	Percent Yes				
	Natural Person	Legal Person	Privacy/ Proxy	Unknown	p-value
On Any Whitelist	14.6	13.4	12.4	3.7	0.0070
Service/Network Providers	9.2	9.0	7.1	1.5	0.0166
No Whitelist Octate	7.0	6.6	5.3	2.2	0.1701

Overall, 12.8 percent of the domains were matched to any whitelist, but this percentage is significantly lower (3.7 percent) for unknown registrant type domains, as shown by a p-value of 0.0070. Similarly, the rate of unknown registrant type domains on a service/network provider whitelist (1.5 percent) is significantly lower than for the other three registrant types (seven to nine percent), as shown by a p-value of 0.0166. The same pattern appears for the no whitelist octate, but the differences are not statistically significant. For all three variables, the privacy/proxy registration rate is slightly lower than registrants who are apparently natural or legal persons.

Potentially Commercial Activity

Table J.5: Domains Found on Whitelists by Potentially Commercial Activity

Whitelist Variable	Percent Yes		
	No Potentially Commercial Activity	Potentially Commercial Activity	p-value
On Any Whitelist	5.6	18.1	<.0001
Service/Network Providers	4.0	11.1	<.0001
No Whitelist Octate	1.9	9.3	<.0001

All three whitelist variables show very statistically significant differences between domains with and without Potentially Commercial activity. Domains with Potentially Commercial activity are much more likely to appear on any whitelist, as well as either of the two whitelist categories with enough positive matches to be separated out (service/network providers and no whitelist octate).

WHOIS REGISITRANT IDENTIFICATION STUDY

Appendix B: Variable Glossary

PRESENTED TO:
ICANN

PRESENTED BY:
NORC at the University of Chicago

MAY 23, 2013

NORC at the UNIVERSITY of CHICAGO

Variable Name	Description	Categories
Allegedly Illegal or Harmful Activity	Behaviors inferred from specific evidence noted by manual coders. Coders received training in the nature of these behaviors and their tell-tale signs on websites.	No allegedly illegal activities detected Spam Advance fee fraud (aka 419 scams) Phishing Cybersquatting/Typosquatting Counterfeit merchandise Trademark infringement Malware Intellectual property theft Child sexual images Identity theft Money laundering
Apparent Domain Registrant Type	For each sampled domain, the type of person who registered the domain as indicated by WHOIS information in "registrant name" and "registrant organization" categories.	Natural Person Legal Person Privacy/Proxy Unknown
Apparent Domain User Type	For each sampled domain, the type of person who appeared to be the beneficial user of the domain. Often inferred from "About" and similar sections on websites.	Natural Person Legal Person Domain Parked No Online Content Unknown

Variable Name	Description	Categories
Blacklist Variables	Codes for allegedly illegal or harmful activity categories as determined by DNS blacklists which received top-rankings for accuracy from NORC analysts. It is possible for a domain to be categorized in more than one way, so the categories are not mutually exclusive.	Abusive Abusive host Abusive host & anonymous-state Backscatter Ddos attacks Dynamic-ip Spam Spam abuse vulnerability Spam bad host, no cookie Suspicious Suspicious & comment spammer Tor network Trojan/virus/bot
Business Structure of Domain User	Based on evidence in domain content of a sampled domain and external databases (e.g. Accurint), the apparent structure of a domain user's business.	For-Profit Business Non-Profit Business Not a Business Unknown Business Structure
Domain Name Extension (gTLD)	The domain name extension of the sampled generic top level domain.	*.com *.net *.org *.info *.biz

Variable Name	Description	Categories
Explicit Sexual Images	Manual determination of whether or not explicit sexual images are present in the web content of a sampled domain	Yes No
Potentially Commercial Activities	The specific type of commercial activities a website appeared to transact, based on web content analysis by manual coders	E-Commerce Membership (Online Content) Membership (Offline Content) Promotional Content (Offline) Promotional Content (Online) Host Promotional Content (Online) Third Party Banner Ads Host Banner Ads Pay-Per-Click Ads Host Pay-Per-Click Ads
Registrant Country/Region Of The World	Based on the WHOIS country of residence for the individual or organization that registered the sampled domain.	United States China United Kingdom Germany Australia/New Zealand Canada Other Europe Other Asia/Pacific Other (North America excluding the U.S. and Canada, South America, Caribbean Islands, and Africa) Ambiguous/Missing
Relationship Of Domain User to Registrant	The discernible relationship between the domain user and domain registrant, often derived by cross-referencing WHOIS information with web content for concordant attributes (e.g. a registrant's role in the domain use is specified on the website)	Domain User Same as Registrant Domain User is Customer of Privacy/Proxy Registrant Domain User is Customer of Other Registrant Domain User is Employer/Employee of Registrant Unable to Determine Relationship

Variable Name	Description	Categories
Whitelist Variables	Codes for domains found on the whitelist hosted by www.dnswl.org and two additional whitelists. The response octate of the dnswl.org gave additional information on the category of the entry on the whitelist. It is possible for a domain to be identified by more than one whitelist, so the categories in the summary table are not mutually exclusive.	Retail/Wholesale Services Service/Network Providers Email Service Providers No Whitelist Octate

WHOIS REGISITRANT IDENTIFICATION STUDY

Appendix C: Report Modifications

PRESENTED TO:
ICANN

PRESENTED BY:
NORC at the University of Chicago

MAY 23, 2013

NORC at the UNIVERSITY of CHICAGO

ICANN released a draft version of this report on February 15, 2013 (the **Draft Report**). Since that time, NORC has modified the report to correct minor errors, and to clarify issues identified in public comments. This appendix provides documentation of the changes so that readers can identify and understand the changes made to the **WHOIS Registrant Identification Study Report** (the **Final Report**).

- References to the *ICANN's Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the Top 5 gTLDs* (the *Privacy/Proxy Prevalence Study*).
  - The **Draft Report** referred to the September 28, 2009 draft release of the *Privacy/Proxy Prevalence Study*, whereas the Final Report references the September 14, 2010 final release of the *Privacy/Proxy Prevalence Study*.
- Section 2.3.1. WHOIS Variables, Privacy/Proxy Services
  - The **Draft Report** used the September 28, 2009 *Privacy/Proxy Prevalence Study* estimate that approximately 24 percent of domains in the top five gTLDs are likely registered using a privacy or proxy service. By the time of the release of the September 14, 2010 *Privacy/Proxy Prevalence Study*, NORC had refined the process for determining when WHOIS registrant information privacy and proxy, and concluded that approximately 18 percent were registered using a privacy or proxy service. The **Final Report** now uses the 18 percent figure from the September 14, 2010 report.
    - In both the **Draft Report** and the **Final Report**, NORC estimates that the percentage of Top Five gTLDs registered using privacy or proxy services is 20 percent. This would be a statistically significant difference with the 24 percent estimate from the September 28, 2009 *Privacy/Proxy Prevalence Study*, but it is <u>not</u> a statistically significant difference with the 18 percent estimate from the September 14, 2010 *Privacy/Proxy Prevalence Study*.
  - In the **Draft Report**, of the 320 domains determined to be registered by a privacy or proxy service, it was stated that 10 of these domains used a privacy service and 310 used a proxy service. During a quality review of the data, NORC determined that 10 of the remaining 310 domains actually used a privacy service. Therefore, the total number of domains registered using a privacy service is 20, and the total number of domain using a proxy service is 300.
    - No additional domains using a privacy or proxy service were identified (out of all domains in the sample). The total number of domains using a privacy or proxy service remains at 320. Therefore, none of the statistics for privacy/proxy services in the remainder of the report changed.
    - Due to the increase in the number of domains using a privacy service, the **Final Report's** estimate of the percentage of domains using a privacy service within the group of 320 domains using a privacy or proxy service increased to 6 percent (from 3 percent in the **Draft Report**).

- The **Draft Report** used the September 28, 2009 *Privacy/Proxy Prevalence Study* estimate that approximately 15 percent of domains registered using a privacy or proxy service were registered using a privacy service. This changed to 9 percent in the September 14, 2010 *Privacy/Proxy Prevalence Study*.
  - The difference between the 9 percent estimate from the September 14, 2010 *Privacy/Proxy Prevalence Study* and the 6 percent estimate in the **Final Report** is not a statistically significant difference.
  - NORC noticed that four privacy/proxy service providers that registered domains found in both the previous *Privacy/Proxy Prevalence Study* and the current **Registrant Identification Study** only provided proxy services for domains in the current study, but provided both privacy and proxy services for domains in the prior study. This is now documented in the **Final Report**.

- Section 2.3.1. WHOIS Variables, Registrant's WHOIS Address Country/Region of the World
  - Because there was interest in the 81 WHOIS records for which no registrant country could be determined, NORC investigated the records further. The **Final Report** now provides more information about these records.
    - For 10 of the sampled domain names, a WHOIS record existed; however, the record did not provide sufficient information to accurately code the registrant's country strictly based on the WHOIS data.
    - For the remaining 71 domain names, there was no WHOIS record—the WHOIS information was completely missing. Because of the dynamic nature of the internet environment it is difficult to determine why this happens. At this point in time we can only speculate as to why these records were missing. It appears that in some cases, the domain expired between the time the domain name sample was selected and the time the NORC-BOT extracted the data. In other cases, a WHOIS record did not exist at the time of the NORC-BOT extraction, but it is possible a WHOIS record may now exist.

- Appendix A, Table 3
  - The table was updated to reflect the changes to the number of registrant organizations that appear to be proxy versus privacy services (see note above).
  - The text following the table was updated in the same way.