WHOIS Registrant Identification Study

ICANN Generic Names Supporting Organization Webinar
March 6, 2013

Edward Mulrow, Project Director
Michael Jugovich, IT Data Leader
Steven Pedlow, Senior Statistician

NORC at the UNIVERSITY of CHICAGO

Outline of Presentation

- Goals of Project
- Sample Design
- Data Collection
- Coding
  - WHOIS variables
  - Domain User variables
  - Domain Content variables
- Answer to GAC Questions
- Lessons Learned

Goals of Project

- Exploratory examination of WHOIS data for a representative sample of top five ICANN gTLDs
- Intent: Understanding Registrants and Domain Users
  - Types of Entities Using these Services
  - Kinds of Activities for these Domains
- Three focus areas
  1. Registrants: Natural Persons, Legal Persons and Privacy/Proxy Services
  2. Domain Users: Natural Persons and Legal Persons
  3. Potentially Commercial Activity: compare Yes and No

Government Advisory Committee (GAC) Questions

- 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 person users?
- What is the relative percentage of Privacy/Proxy use among domains with commercial use?

Sample Design

- Stratification only possible by gTLD
- Scope limited to five most common gTLDs (98.5 Percent*)

gTLD	Global Percentage	Sample Size	Sample Percentage
*.com	74.3	1,128	70.5
*.net	10.7	165	10.3
*.org	7.2	107	6.7
*.info	6.1	100	6.3
*.biz	1.6	100	6.3
TOTAL	100.0	1,600	100.0

* Based on June 2011 Registry Operator Monthly Reports
http://www.icann.org/en/tlds/monthly-reports/

Data Collection

- NORC-BOT is a multi-threaded application (Python 2.7)
- NORC-BOT automated information gathering tool
  1. WHOIS data
  2. Publicly accessible HTTP/HTTPS/FTP files
  3. Response codes from DNS Blacklists (including Whitelists)
- This information is not static, but a point in time
- We attempted to simultaneously extract all three
  - WHOIS data obtained from the WHOISAPI service
  - WHOISAPI did not always return WHOIS information
  - ICANN ran own WHOIS extraction in parallel
  - Two simultaneous extractions merged and compared
- Data collection was complete in March 2012

Web/FTP Content Extraction

- Only the *www* and *ww2* subdomains were searched for content for each sampled domain
- Domains may have had content on other subdomains, and no attempt was made to look for such content
- A download quota (100MB) was necessary to ensure that extremely large sites hosting GBs of content were not indexed

Coding – Three Broad Classes of Variables

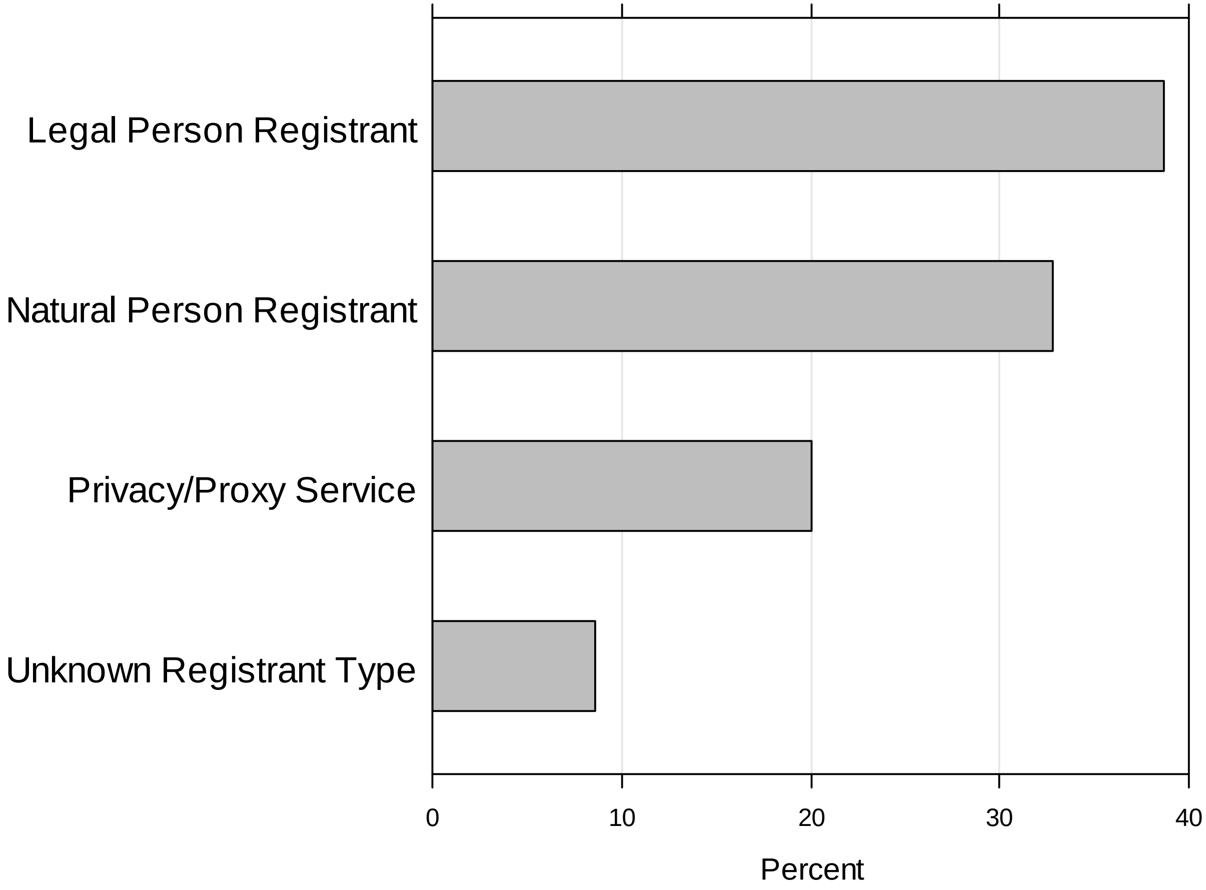
- WHOIS
  - **Apparent Registrant Type**, Country/Region of the World, Registrar
  - Coded based on WHOIS information and independent searches of public databases
- Domain User
  - **Apparent Domain User Type**, User Relationship to Registrant, User Business Structure
  - Coded based on downloaded content
- Domain Content
  - **Potentially Commercial Activity**, Allegedly Illegal or Harmful Activities, Explicit Sexual Imagery
  - Coded based on downloaded content

Coding – Apparent Registrant Type (WHOIS)

- Natural Person: WHOIS data appeared to identify a real living individual
- Legal Person: WHOIS data appeared to identify a company, business, partnership, non-profit entity, trade association, etc.
  - Includes multiple domain holders, but not Privacy/Proxy service providers
    - reverse WHOIS email counts were used to help determine multiple domain name holders
- Privacy/Proxy Service: WHOIS data appeared to identify a Privacy/Proxy service
  - Used lists of known providers constructed for the "*Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service*" as a guide
- Unclassified: classification was not apparent based on WHOIS data
  - Includes data completely missing, patently false, or incomplete, and domains pending reactivation or deletion

GAC Question: Apparent Registrant Type

What is the percentage of registrants that are natural versus legal persons?



39% ± 2%

33% ± 2%

20% ± 2%

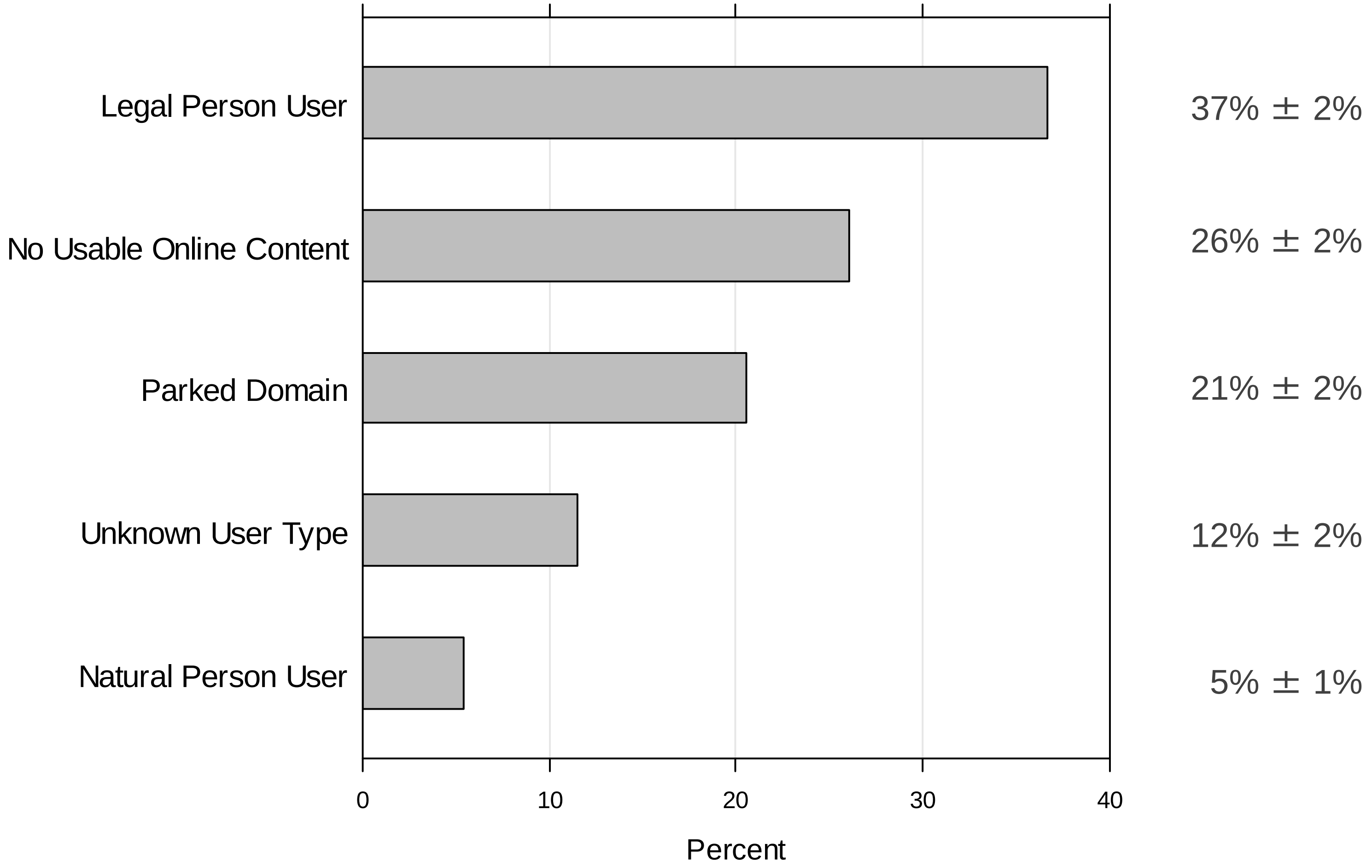
8% ± 1%

All percentages are based on the complete set of 1,600 sampled domains

Coding – Apparent Domain User Type (Domain User)

- Natural Person – domain content appeared to identify the domain user as a real living individual
- Legal Person – domain content appeared to identify the domain user a company, business, partnership, non-profit entity, trade association, etc.
- No Usable Online Content – no content available, or minimal HTML code existed but it was insufficient to determine the user type
- Parked Domain – similar to No Usable Online Content, but the domain landing page's minimal HTML content was consistent with typical domain parking content
- Unknown User Type – available content, but NORC could not determine the user type (natural or legal person)

Apparent Domain User Type

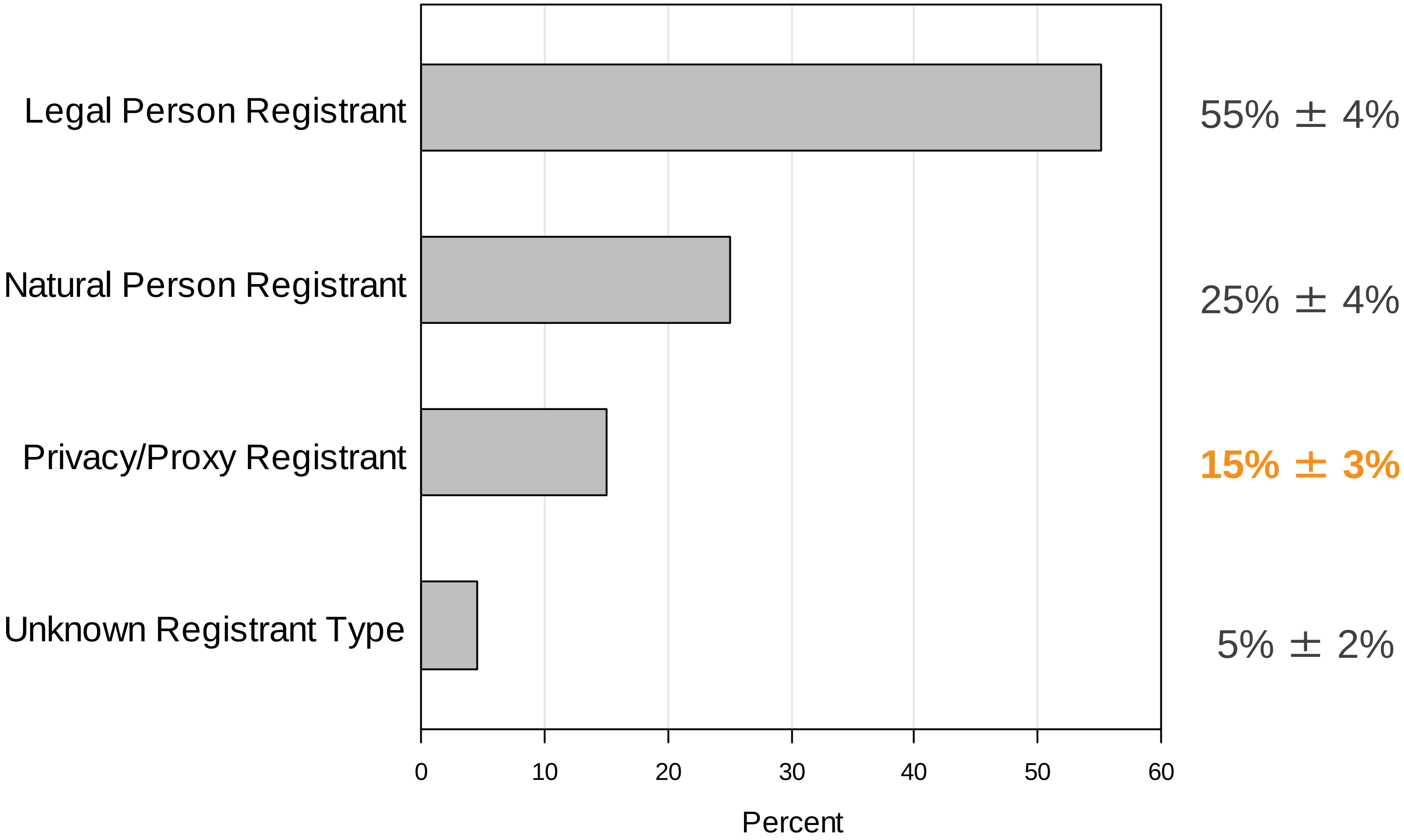


NORC at the UNIVERSITY of CHICAGO

GAC Question: Privacy/Proxy Relative to Legal Person Domain Users

What is the relative percentage of Privacy/Proxy use among legal person users?

Apparent Registrant Type Relative to Legal Person Users



All percentages are based on 586 sampled domains with Domain Users classified as Legal Persons

Coding – Potentially Commercial Activity (Domain Content)

- Attempted to categorize all observed monetary activities that in some countries might be legally considered “commercial activities”
- Looked for evidence of
  - e-commerce
  - collection of membership dues for online or offline content
  - promotional material content
  - banner ads
  - pay-per-click ads

Potentially Commercial Activity (Domain Content)

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

* A domain could show evidence of one or more of the activities

GAC Question: Potentially Commercial Activity (PCA)

What is the percentage of domain name uses that are commercial versus non-commercial?

When Pay-per-click Ads are considered Potentially Commercial Activity

- At least one of the five activities was detected in 905 of the 1,600 sampled domains
  - **57% ± 2%**

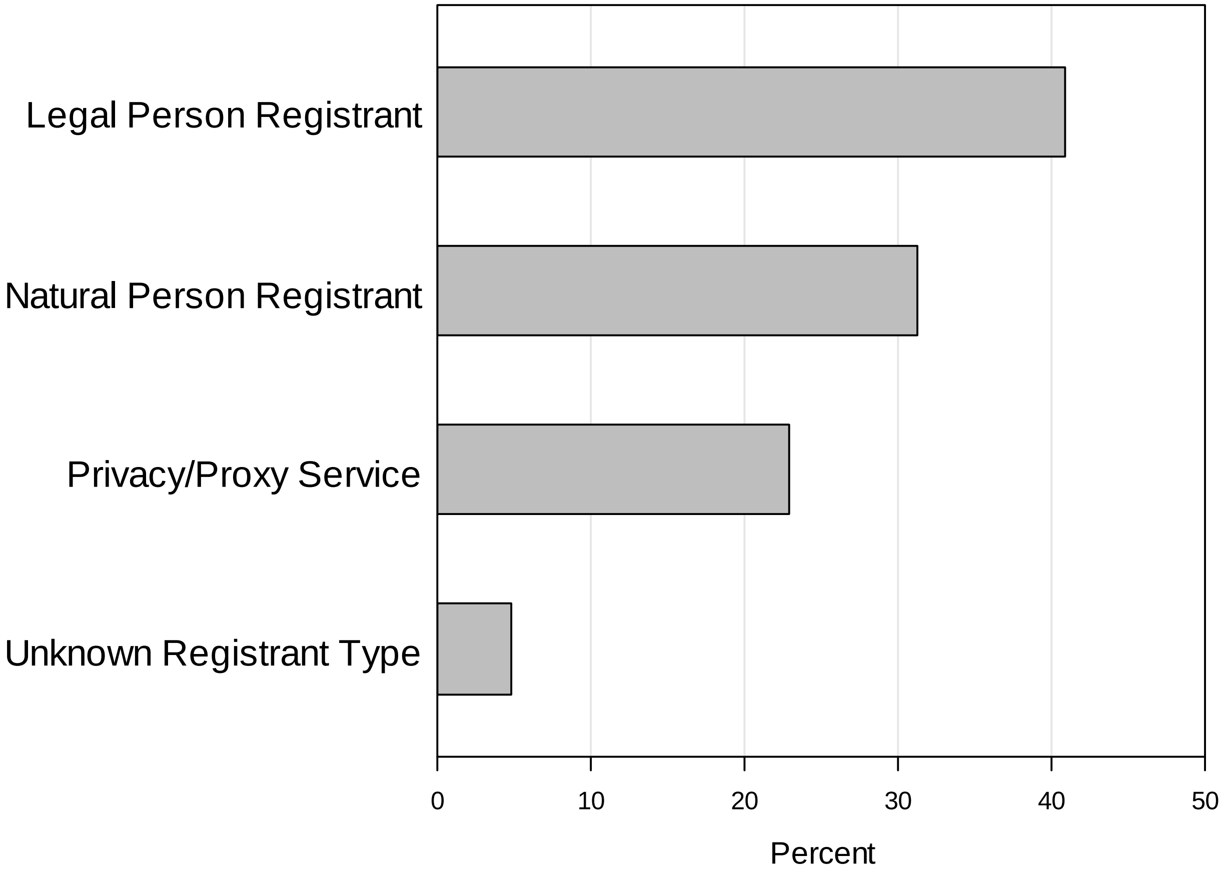
When Pay-per-click Ads are not considered Potentially Commercial Activity

- If pay-per-click ads are not considered PCA, the number of domains with at least one activity drops to 717
  - **45% ± 2%**

GAC Question: Privacy/Proxy Relative to Potentially Commercial Activity (PCA)

What is the relative percentage of Privacy/Proxy use among domains with commercial use?

Apparent Registrant Type Relative to Domains with PCA



41% ± 3%

31% ± 3%

23% ± 3%

5% ± 1%

All percentages are based on 905 sampled domains with detected Potentially Commercial Activity (includes Pay-per-click Ads)

Lessons Learned

- Collecting WHOIS, domain content, and DNSBL information in a nearly simultaneous manner is difficult, but a multi-threaded application such as NORC-BOT makes the task feasible
  - NORC's Registrant ID Summary Report contains a summary of NORC-BOT features
- Data coding provided subjective challenges due to the inherent ambiguity of internet data
- Attempt to impose standard codes on a huge variety of unique websites revealed the "fuzziness" of some prevailing concepts used in studying Internet activity
  - For example, domain parking versus domain reselling
- Domain User Relationship to the Registrant was difficult to determine (55% Unknown, but related to domains without content)
- Domain User's Business Structure was also difficult to discern (65% Unknown), and contrary to our hypothesis, did not provide additional insight into the registrant/user relationship

Summary

- Exploratory study is a first step in ICANN's process to learn about domain name registrants and their relationships to domain users and the ways in which domains are used
- In many cases, classification of the characteristics and activities were difficult to discern and often had to be coded as "unknown."
- A large enough number of domains were able to be coded so that important relationships were uncovered.
- NORC's draft report is available at

http://www.icann.org/en/news/public-comment/whois-regid-15feb13-en.htm

- ICANN is seeking comments from the public which can be submitted at the same site
    - Comment period closes on March 31, 2013

Thank You!

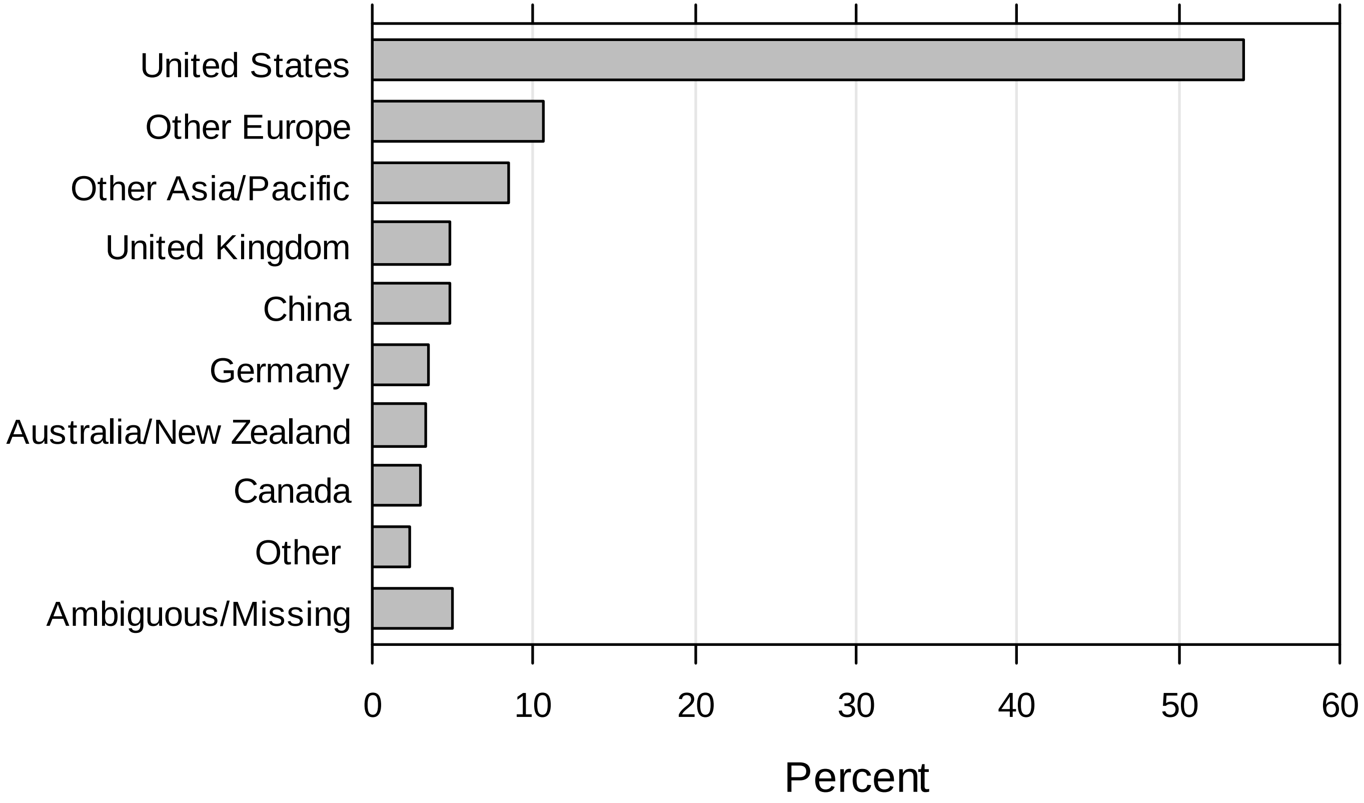
NORC at the UNIVERSITY of CHICAGO

Additional supporting material follows. This material also appears in the Registrant Identification Study Draft Report.

http://www.icann.org/en/news/public-comment/whois-regid-15feb13-en.htm

insight for informed decisions™

Registrant’s Address Country/Region of the World (WHOIS)

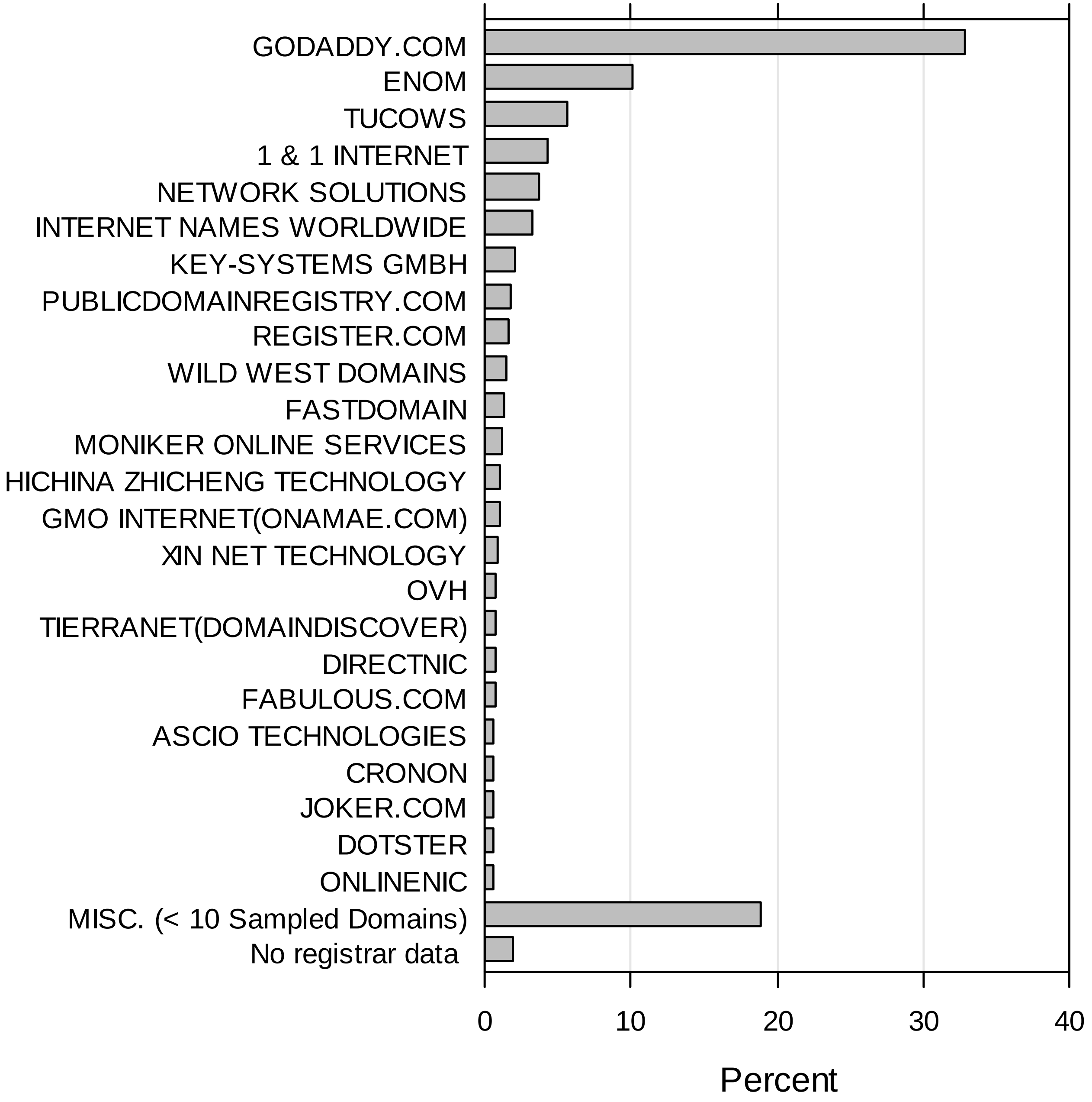


Other Europe = European countries other than the U.K. or Germany;

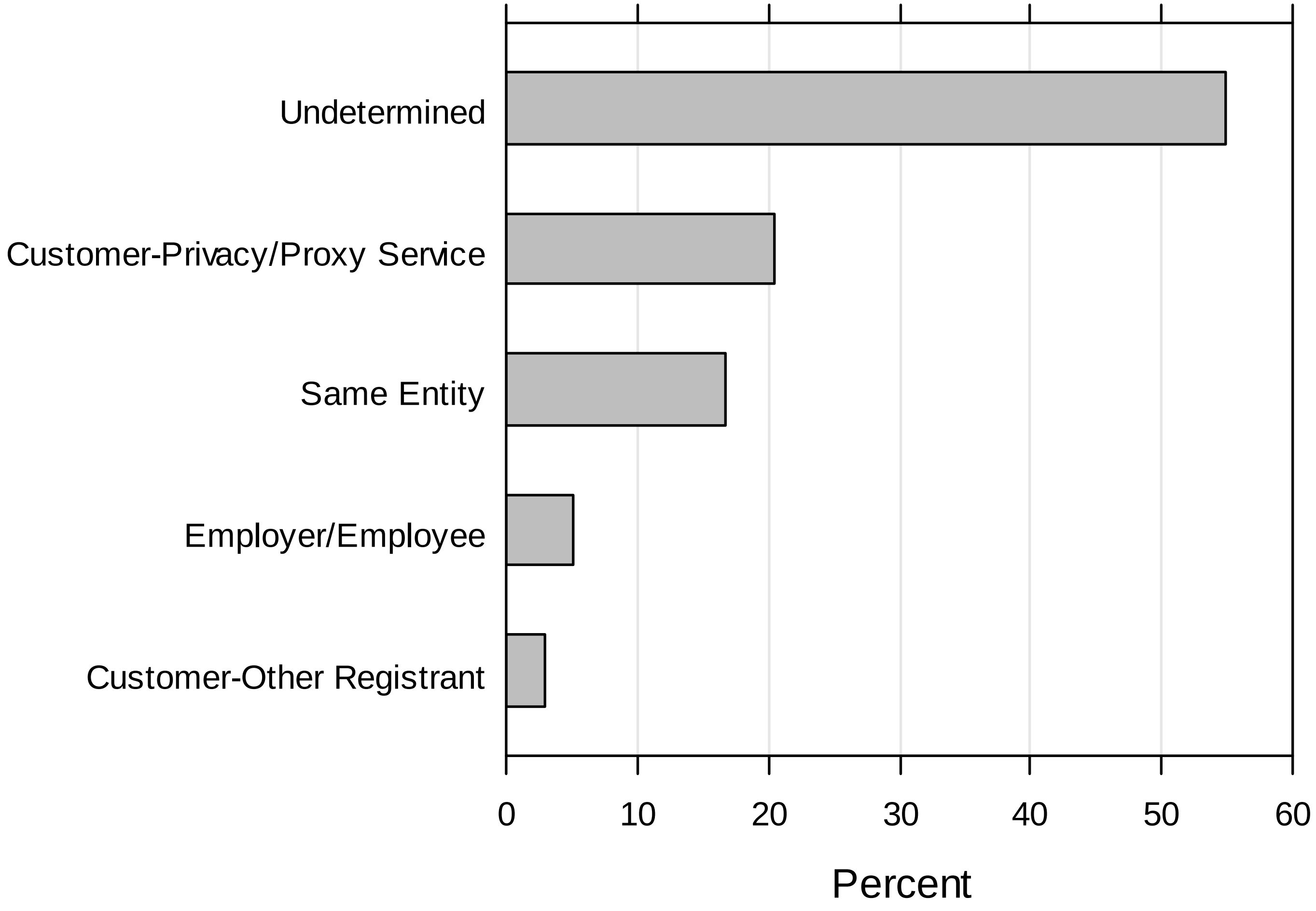
Other Asia/Pacific = Asian/Pacific countries other than China, Australia, or New Zealand

Other = countries in any of the following regions: North America excluding the U.S. and Canada, South America, Caribbean Islands, and Africa

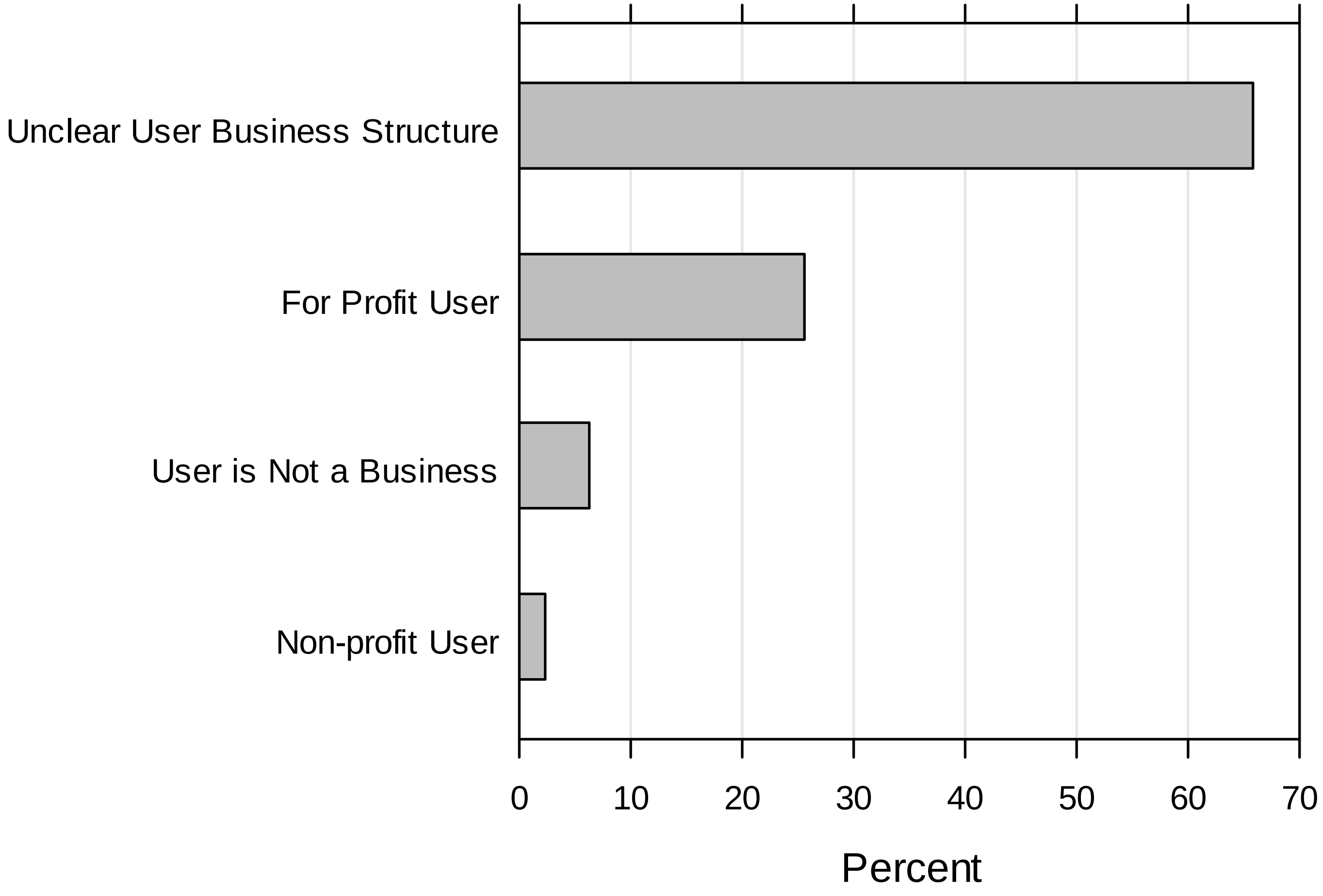
WHOIS – Domain Registrars



User Relationship to Registrant (Domain User)

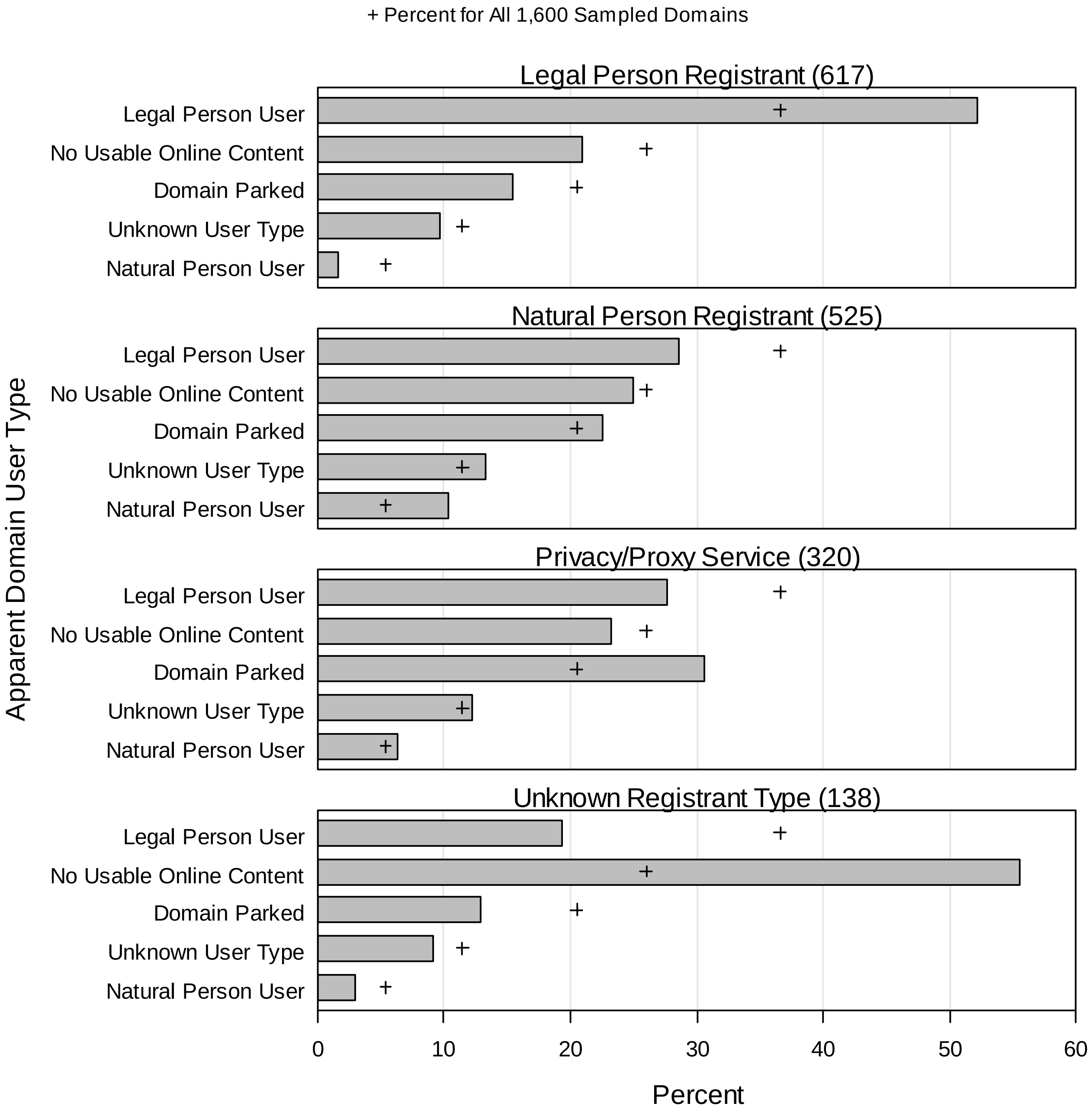


Domain User – Apparent Business Structure

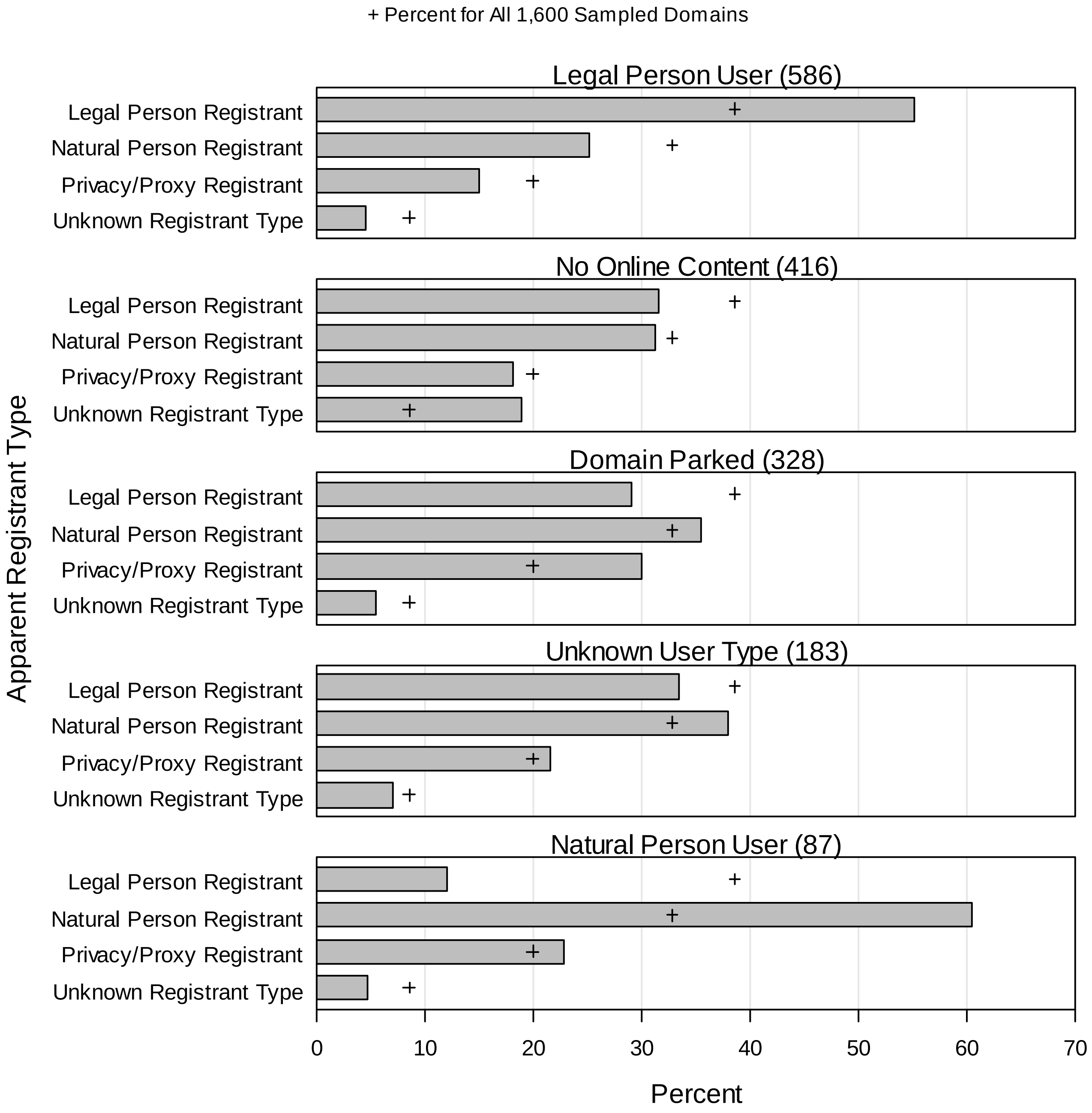


- Determined through manual inspection and keyword searches
  - Consulted third-party databases
    - **Accurint, LinkedIn, digitalenterprise.org/models/models.html**

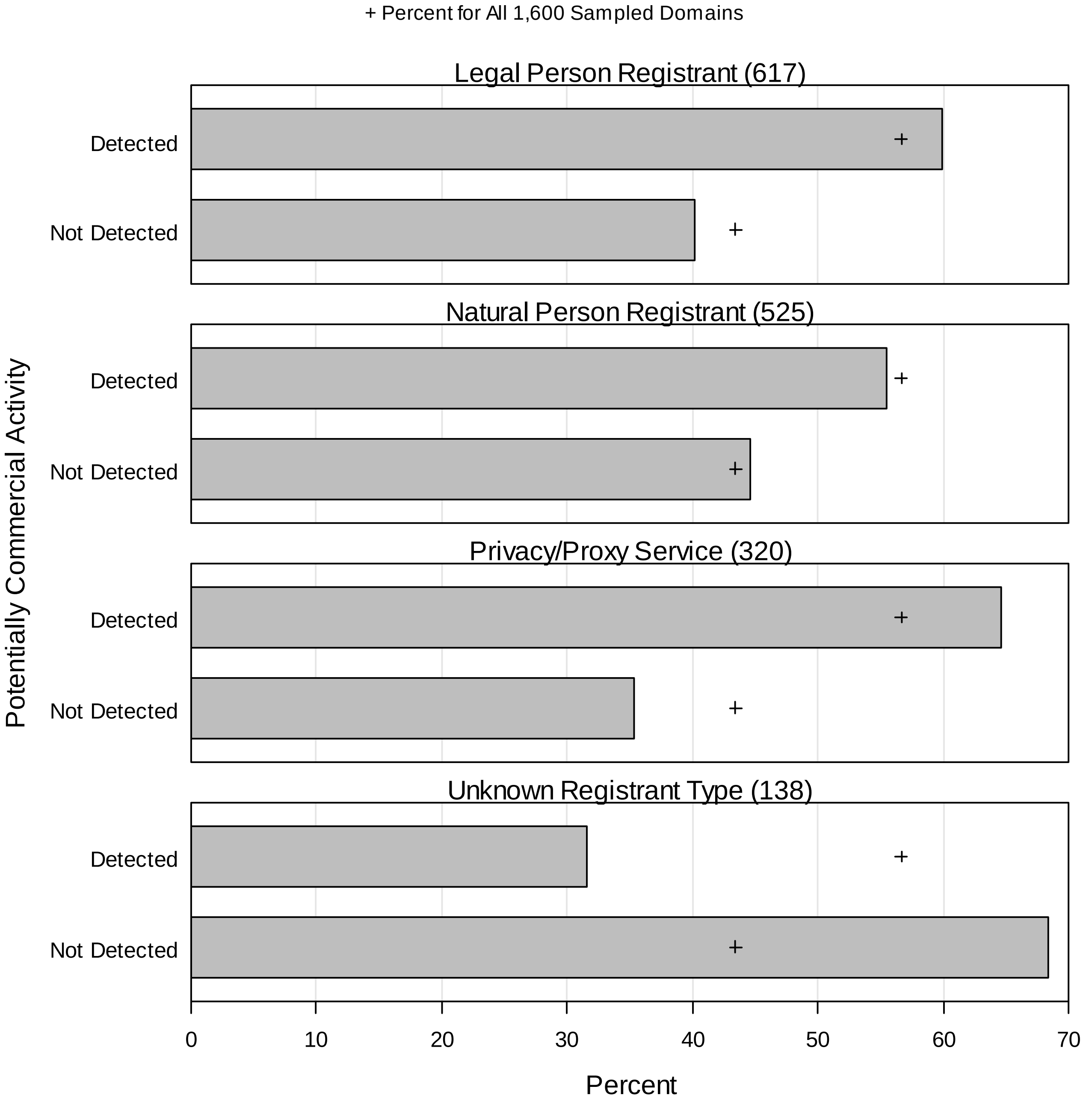
Domain Use by Registrant Type – Domain User Type



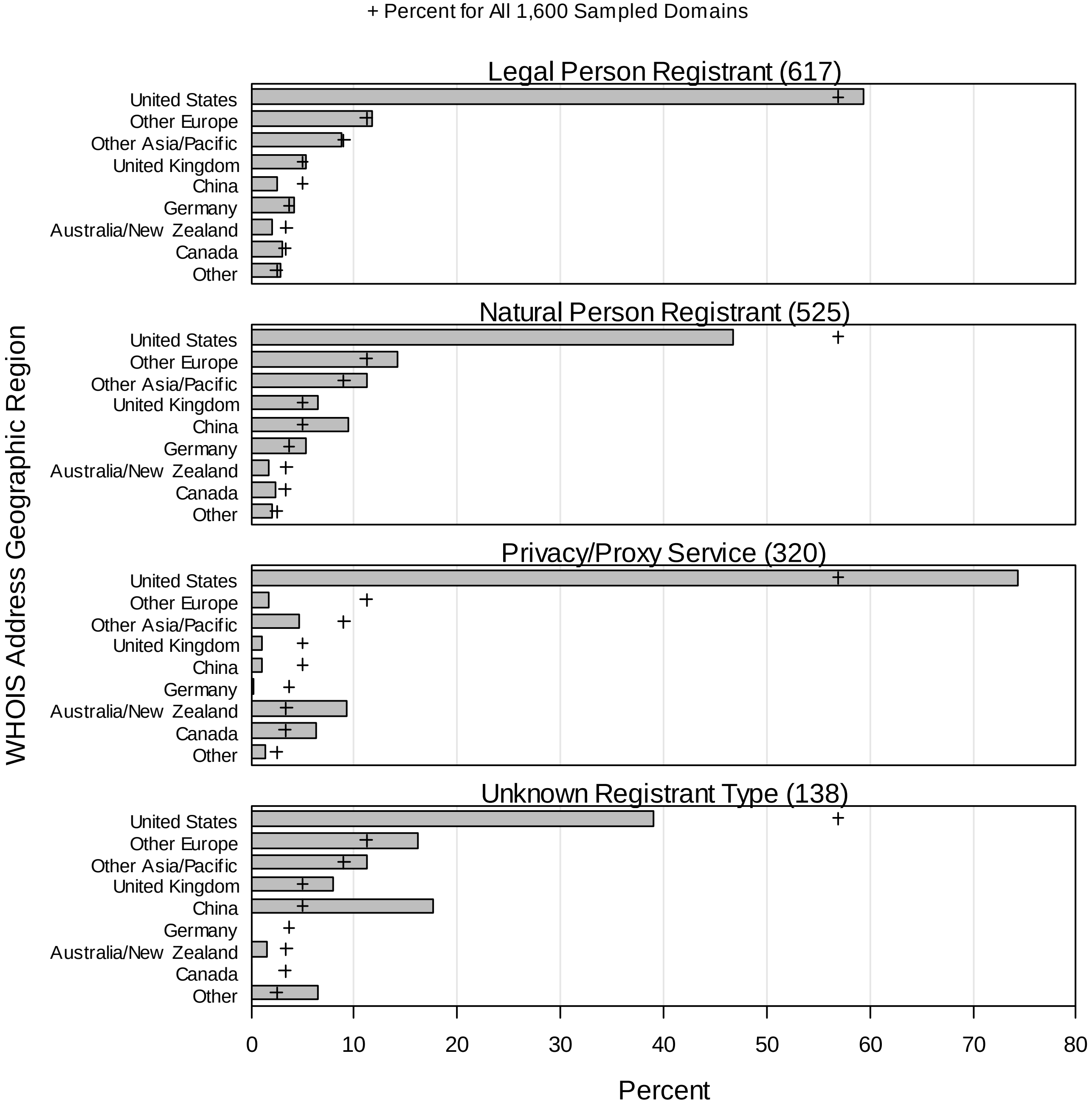
Registrant Type by Domain User Type



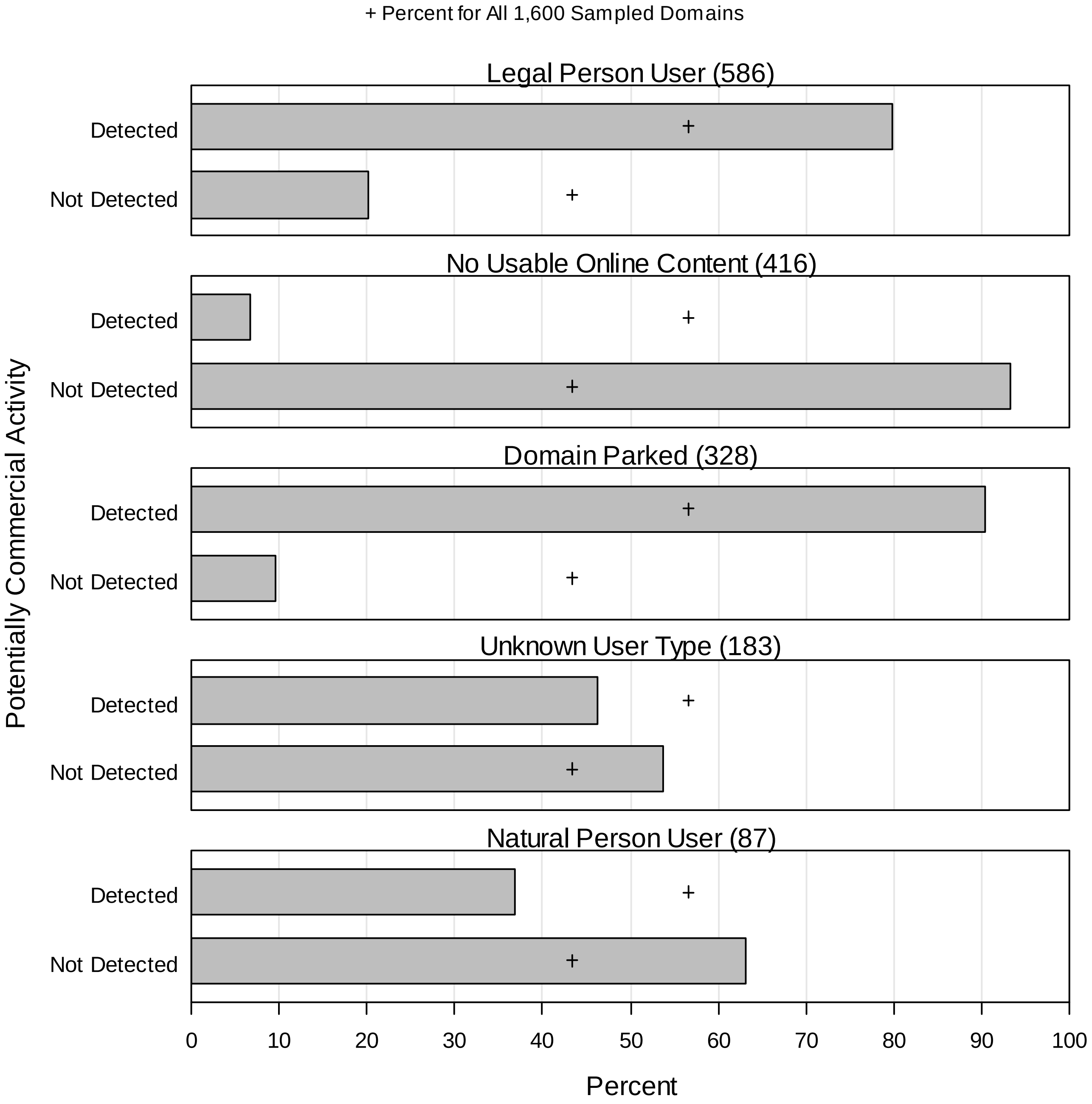
Domain Use by Registrant Type – Potentially Commercial Activity



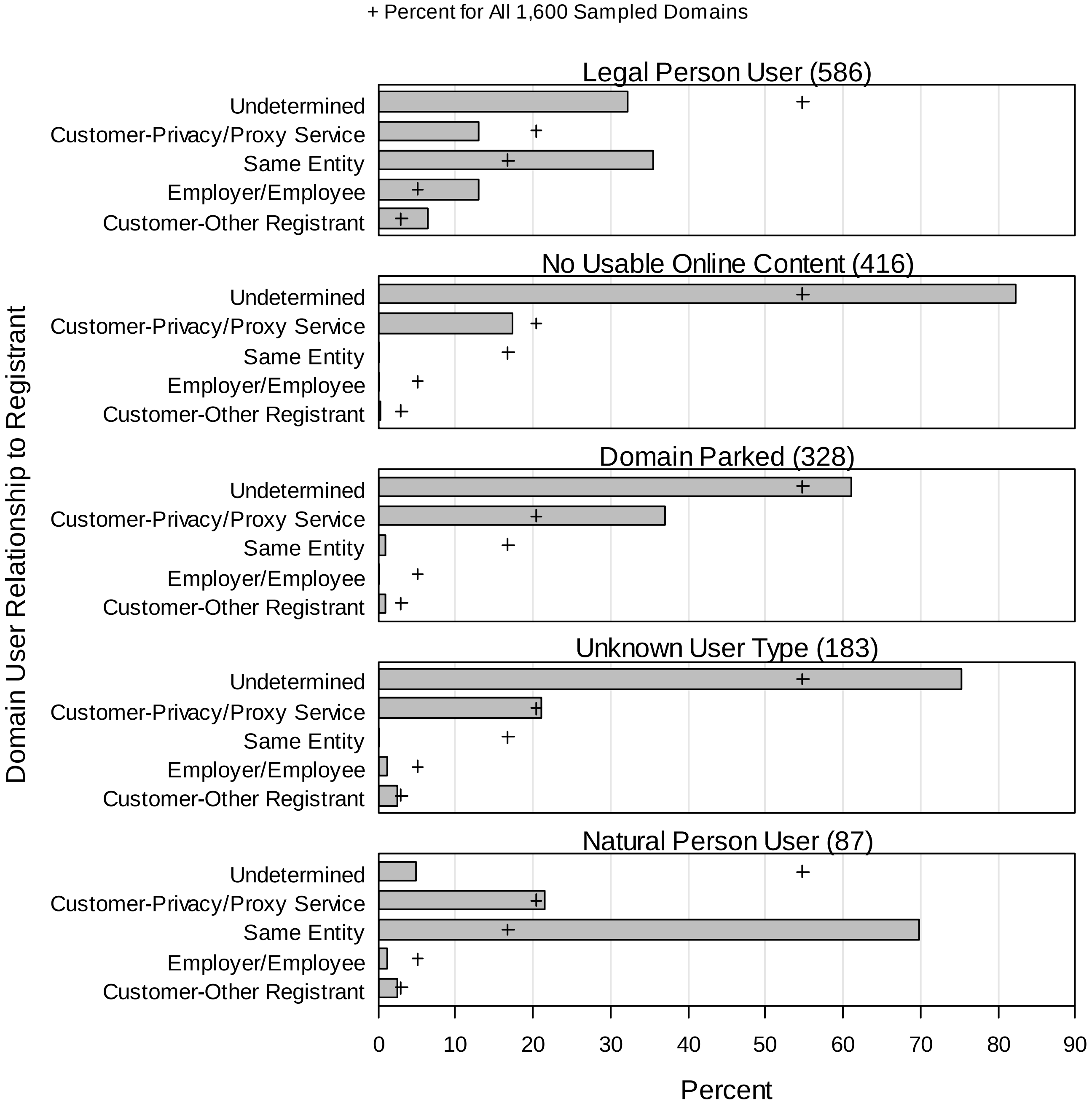
WHOIS Country/Region by Registrant Type



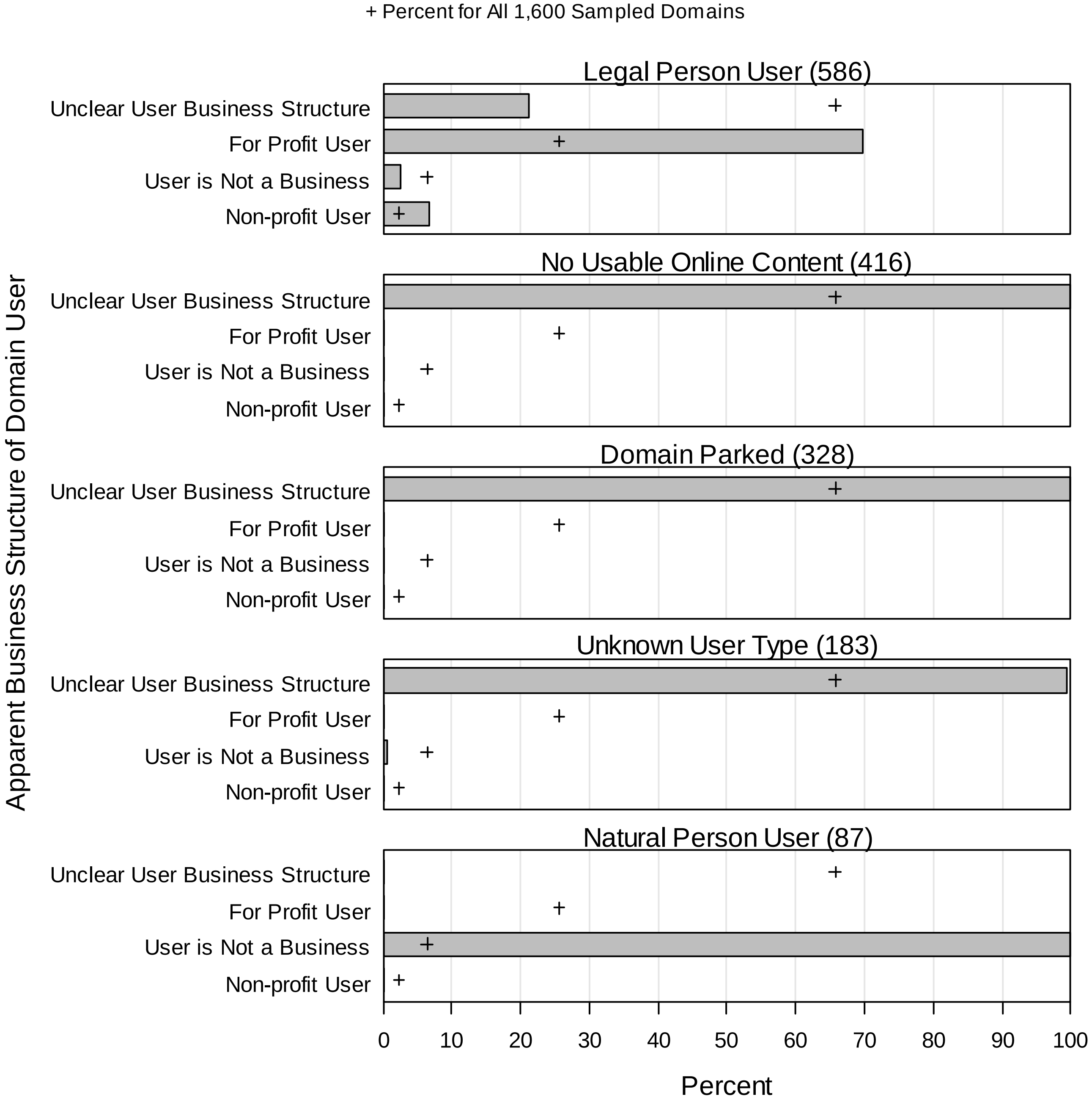
Domain Use by User Type – Potentially Commercial Activity



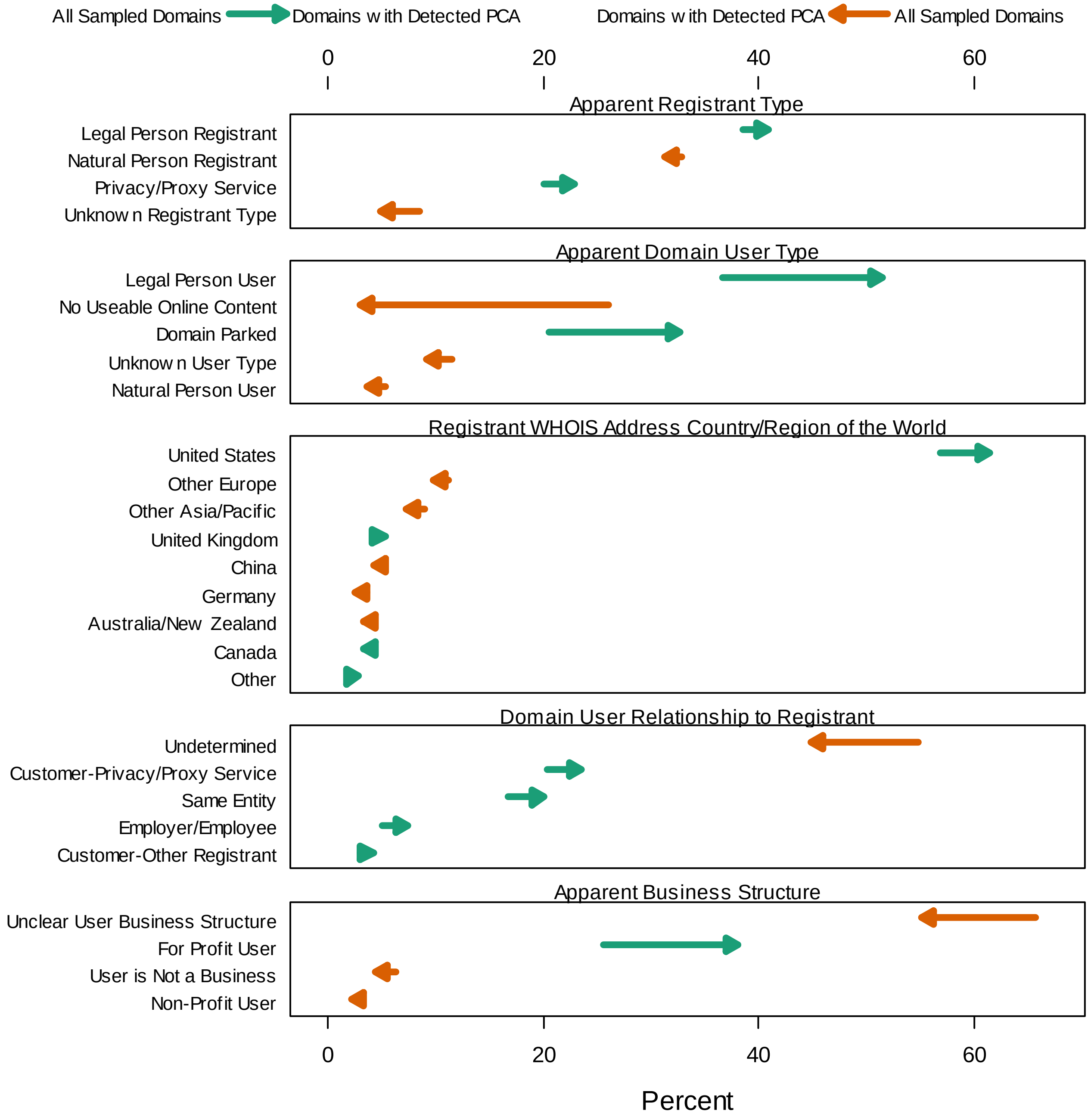
Domain Use by User Type – User/Registrant Relationship



Domain Use by User Type – Business Structure



Domains with Potentially Commercial Activity (PCA) versus Overall



Domain Content – Allegedly Illegal or Harmful Activities and Explicit Sexual Content

- Manually coded based on coders judgment of observed HTTP/FTP domain content.
- Automated coding based on DNS Blacklist scans
  - Blacklists: lists of IP addresses of computers or networks linked to allegedly illegal or potentially harmful activities
  - Whitelists: lists used to exempt a domain or URL from black-listing

Allegedly Illegal or Harmful Activities and Explicit Sexual Content – Manually Coded

- Only 18 domains were manually classified as having allegedly illegal or harmful activities (1.1 percent)
- Only 16 domains were observed to contain explicit sexual content (1.0 percent)
- Further cross-classified analysis of these data for the purpose of determining if these two behaviors are more likely among certain subgroups is questionable given the small number of observations
  - ICANN has commissioned a separate study to explore privacy/proxy abuse. It is exclusively focused on finding domains engaged in allegedly illegal or harmful activity

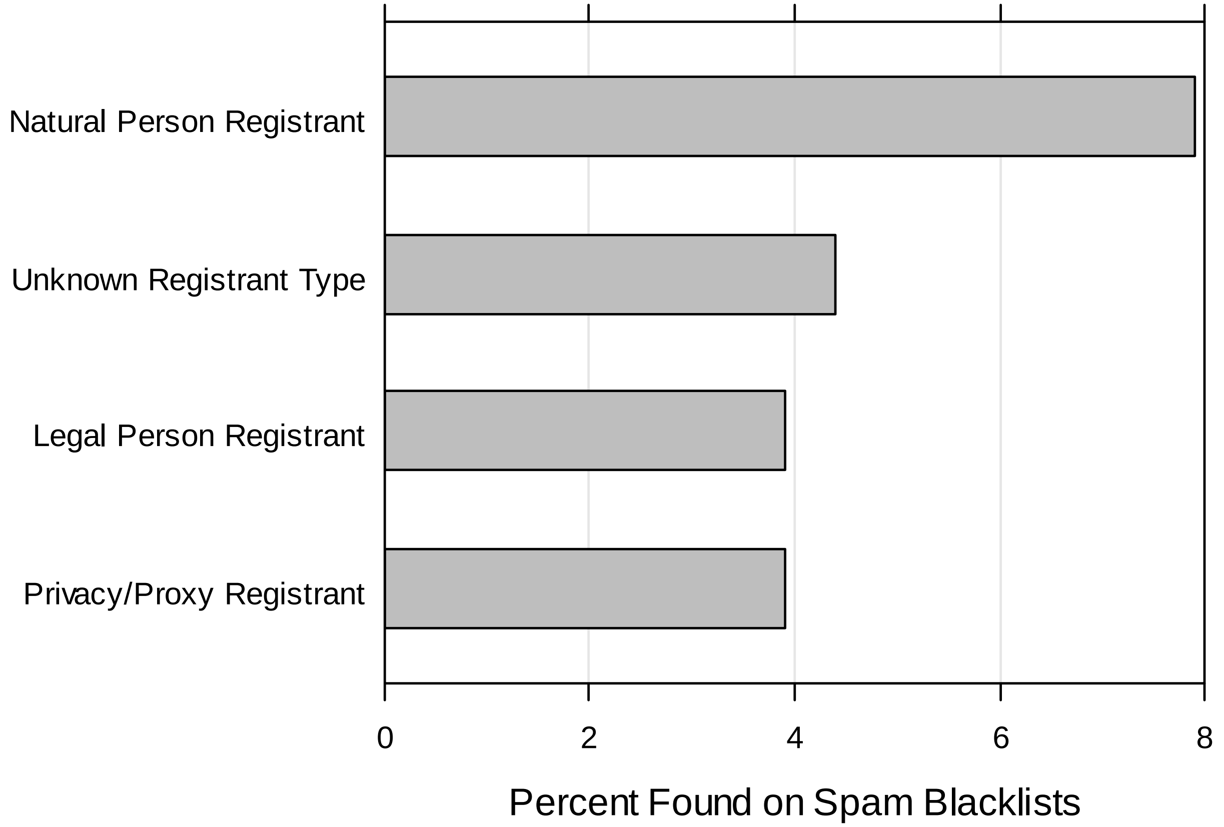
Allegedly Illegal or Harmful Activities – Blacklist Scans

- 141 sampled domains were found on at least one blacklist (8.8 percent).
- 204 sampled domains were found on whitelists (12.8 percent)
  - Whitelists are lists used to exempt a domain or URL from black-listing
  - 13 of these were also found on a blacklist

Allegedly Illegal or Harmful Activities Blacklist Scans

- Cross-classified results 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 do; especially for spam monitoring blacklists

Presence on Spam Blacklists Within 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.