Contents

WHOIS Proxy / Privacy Service Abuse Study –Definition

This study will measure how often domains associated with illegal or harmful Internet communication abuse Privacy/Proxy services to obscure the perpetrator's identity.

1. Objective

This study is intended to help the ICANN community determine the extent to which Proxy and Privacy services are abused during illegal or harmful Internet communication. Specifically, it will attempt to prove/disprove the following hypothesis:

> **A significant percentage of the domain names used to conduct illegal or harmful Internet activities are registered via Privacy or Proxy services to obscure the perpetrator's identity.**

As defined by [1], "illegal or harmful communication" refers to online activities (e.g., email messages, web transactions, file downloads) that violate criminal or civil law or which harm their targets (e.g., email/download recipients, website visitors). These activities include unsolicited commercial bulk email (spam), online intellectual property or identity theft, email harassment or stalking, phishing websites, online malware dissemination, and cybersquatting. Further examples include DoS attacks, DNS cache poisoning, pirated software (warez) distribution sites, money laundering email (mules scams), advanced fee fraud email (411 scams), and online sale of counterfeit merchandise or pharmaceuticals.

Allegations of actionable harm may require victims, law enforcement officials, and others to contact domain users (i.e., owners or licensees). To facilitate identification and contact, section 3.3.1 of the ICANN Registrar Accreditation Agreement (RAA) [4] requires Registrars to provide an interactive web page and a port 43 WHOIS service to enable free access to up-to-date data concerning all active registered domain names. This WHOIS data includes the name and postal address of the Registered Name Holder and technical and administrative contacts for the domain.

According to [1], Proxy and Privacy registration services provide anonymity or privacy protection for domain users. *Privacy* services hide certain user details from WHOIS by offering alternate contact information and mail forwarding services while not actually shielding the user's identity. *Proxy* services have a third-party register domain names on the user's behalf and then license the use of the domain name so that a third-party's contact information (and not the licensee's) is published in WHOIS. According to the WHOIS Privacy/Proxy Prevalence Study [3], approximately 15 to 25 percent of gTLD domain names are likely to be registered using a Privacy or Proxy service.

Study proposals [8][9][10] suggest that Privacy/Proxy services are being abused to obscure the identity of perpetrators that instigate illegal or harmful Internet communication, thereby impeding investigation. For example, proposal [8] indicates that Privacy/Proxy registrations lengthen phishing website take-down times. Proposal [9] indicates that Privacy/Proxy services are being abused to shield cyber squatters (i.e., parties that register or use a domain name in bad faith to profit from someone else's trademark).

A recent study of 384 domains hosted by ISP 3FN (shut down in June 2009 for abetting criminal activity) found that 38 percent were registered to Proxy services [11]. Of those, approximately half were associated with least one kind of illegal activity. Although small and informal, this study illustrated that domains used by criminals do use Proxy services – in this case, more often than the random domains studied by [3].

To provide the ICANN community with empirical data to evaluate such concerns, this study will methodically analyze a large, broad sample of domains associated with various kinds of illegal or harmful Internet activities. It will measure how often these alleged "bad actors" abuse Privacy/Proxy services, comparing rates for each kind of activity to overall Privacy/Proxy rates measured by [3]. If those rates are found to be significant, policy changes may be warranted to deter Privacy/Proxy abuse.

Note: This study will NOT measure the frequency of illegal/harmful Internet activity. This study will gather a representative sample of illegal/harmful incidents to measure how often Privacy/Proxy services are abused by perpetrators (alleged and confirmed).

2. Approach

This hypothesis will be tested by performing a descriptive study on a representative sample of domains within the top five gTLDs (.biz, .com, .info, .net, .org). To focus on study goals, this sample will be composed exclusively of domains involved in illegal or harmful Internet communication, as documented by organizations that routinely track, investigate, and/or remediate various kinds of activities. To measure frequency of abuse, this study will divvy sampled domain users into those that can be reached directly using WHOIS data and those that must be contacted via a referenced Privacy/Proxy service.

Because creating a single sample that proportionally represents every major kind of illegal or harmful Internet communication is unrealistic, subsamples will be created for each activity to be studied (e.g., a spam sender list, a warez site list). Many domains are likely to be associated with multiple activities and may thus appear in more than one subsample. However, rates will be measured independently for each subsample to determine which activities most often abuse Privacy/Proxy services.

Furthermore, because the nature and duration of illegal/harmful Internet activities varies, different methods will be required for incident tracking, investigation, and remediation.

- Timely response is essential for extremely **short-lived activities** (e.g., spam, phishing, DoS attacks). Where possible, domain subsamples for these activities will be generated by monitoring **live-feeds** (e.g., real-time blacklists), letting researchers query and record WHOIS data in near-real-time.

- Timely response is less critical for activities associated with **long-lived activities** (e.g., trademark infringement, cybersquatting). Subsamples for these activities would be impossible to generate in near-real-time; live-feeds do not exist. Instead, these domains and WHOIS data will be **recorded over time** by study participants routinely involved in these incidents (e.g., first responders and real-time cybercrime researchers, complaint centers and law enforcement agencies, victim advocates).

To meet this study's goals, Privacy/Proxy determination must be based on WHOIS data as it was at the time of the incident. WHOIS queries usually return Registrant data long after an offending domain's web, file, or mail servers disappear, appear on an RBL, or are taken down. However, WHOIS data may well change following illegal activity, such as when a malicious domain is suspended or re-registered. Study goals can still be met so long as a significant percentage of WHOIS queries performed shortly after incidents do not return recently-updated or no Registrant data.

Note that other WHOIS studies [3][6][7] have been defined to measure the overall frequency of Privacy/Proxy use, what types of entities (e.g., natural or legal persons) commonly use Privacy/Proxy-registered domains and for what apparent purpose (e.g., personal or commercial), and how Privacy/Proxy providers respond to domain user reveal requests. Those questions are therefore outside the scope of this study.

However, overall frequency of Privacy/Proxy use [3] must be considered when sizing this study's subsamples so that they represent the top 5 gTLD domain population with a 95% confidence interval. Furthermore, because harmful/illegal Internet communication tends to originate from certain countries and regions, live-feeds and incident reports may be geographically skewed. To reflect world-wide experiences, subsamples must be generated from input sources with international scope – for example, global RBLs.

Finally, this study should build upon the foundation laid by the WHOIS Accuracy Study [2] and WHOIS Privacy/Proxy Prevalence Study [3] as follows.

- **Sample Cleaning and Coding:** WHOIS data for every domain name must include certain mandatory values (e.g., Registrant Name), but there is no RFC-standard record format or even a single global database from which WHOIS data can be obtained. The Accuracy Study [2] developed a methodology for cleaning sampled domain WHOIS data to eliminate parsing errors, translate non-ASCII characters, map Registrants to country code/name, and sort the sample by Regional Internet Registry.

- **Registrant Type Classification:** Next, based on WHOIS Registrant Name and Organization values, the Accuracy Study assigned each sampled domain one of the following Apparent Registrant Types: name completely missing or patently false, a natural person, an organization with or without a person's name, a multiple domain name holder (ISP or reseller), or a potential Privacy/Proxy service provider. All potential Privacy/Proxy service providers were then either confirmed or reclassified.

Even though this study's sample design process and parameters differ, researchers are strongly encouraged to apply the same sample cleaning, coding, and classification process to reduce cost and promote consistency across all WHOIS studies. In particular, the Accuracy Study's methodology for confirming potential Privacy/Proxy use should be applied, as this is the key differentiator upon which this study's findings will be based.

3. Inputs

The first step in conducting this study will be to generate subsamples of domain names associated with each kind of illegal or harmful Internet communication to be measured. As noted in Section 2, because activity nature and duration varies, this study will employ two different research methods: Live-Feed Monitoring for incidents typically reported in real-time and Offline Third-Party Recording for all other kinds of incidents.

Method 1: Live-Feed Monitoring

Domain names associated with the following short-lived illegal/harmful Internet activities should ideally be collected from live-feed sources. Possible sources are listed below; additional suggestions are welcome. Researchers are expected to refine and finalize this source list during the first phase of the study.

As alleged "bad actors" are identified from live-feeds, reverse DNS lookups and WHOIS queries will be performed in near-real-time[1] to record the Registrant' Name, Organization, and Address for domain names associated with each incident. Note that "associated domain name" depends upon the type of activity (e.g., spam sender, phishing website, malware server).

Note that, after incident investigation, many alleged bad actors do not end up being the real perpetrators. For example, many spam senders and phishing servers will be "bots" -- compromised hosts used by criminals without the Registrant's knowledge. Furthermore, domains may be added to RBLs based on complaints rather than verified incidents.

However, these "false positive" incident reports still require investigation; WHOIS Registrant data for those domains plays a role in enabling (or inhibiting) investigation. Therefore, this study must gather and analyze the WHOIS data associated with *all* alleged bad actors (proven or otherwise). To avoid skewing results, this study will *also* analyze refined samples that have been filtered to weed out low-probability cases – for example,

[1] Researchers will need to work around port 43 rate limits by pacing WHOIS queries, retrying failed queries, arranging for preferential access from a WHOIS query provider, or enlisting the help of a live-feed supplier that already has preferential access.

eliminating domains associated with fewer than N reported incidents. Objective sample filtering methods should be defined by researchers at study start; suggestions are welcome.

Once sufficiently large subsamples have been collected for each activity, they will be cleaned, coded, and classified by Registrant Type as described in Section 2 for statistical analysis as described in Section 4.

- **Spam:** Live-feeds from several major real-time Domain Name System Blacklists (DNSBLs) could be used to generate a subsample of spam sender IP addresses/ranges and associated unique domain names. Possible sources include Spamhaus Blocklist, Mailshell Live-Feed, SURBL, URIBL, and SORBS DNSBL.

- **Phishing:** Several major Phishing website live-feeds could be used to generate a subsample of phishing URLs and the domain names that host them. Possible sources include OpenDNS PhishTank and Internet Identity RealPhish.

- **Malware:** A subsample of domains used to host and disseminate malware could be created from live-feeds maintained by major malware researchers and/or Internet security vendors. Possible sources include SRI Malware Threat Center, FireEye Malware Analysis & Exchange, and Malware Domains.

- **Denial-of-Service and DNS Cache Poisoning:** Input is requested on live-feed sources that could be used to generate subsamples of domains that send harmful messages during these time-sensitive attacks. Potential sources include the IMPACT Global Response Centre NEWS feed and FIRST-member incident response teams.

Method 2: Offline Third-Party Recording

Domain names associated with less time-critical illegal/harmful activities will be gathered from third-parties that routinely respond to or track such incidents in large volume and might be willing to assist by recording WHOIS data early in their investigation. Candidates include first responders and real-time cybercrime researchers, Internet crime complaint centers and law enforcement agencies, and victim advocates. Possible participants are listed below; additional suggestions are welcome. Researchers are expected to refine and finalize this participant list during the first phase of the study.

Consistency and accuracy of reported data is always a concern whenever numerous independent parties supply input for aggregate statistical analysis. To address this concern, researchers will develop a short, simple incident reporting form and process that participants can use to record the type of illegal/harmful activity, associated domain name, and WHOIS Registrant Name, Organization, and Address in a timely fashion. Here again, note that "associated domain name" depends upon the type of activity (e.g., phishing website, warez server, money laundering email sender).

At study start, researchers will identify and invite representative sources to participate. All participants must agree to record and report all incidents encountered as part of their normal operation during a specified study period (e.g., 30 days). In particular, participants shall be asked to report all alleged perpetrators (proven or otherwise), and to indicate whether investigation confirmed or refuted their alleged involvement in the incident. This data collection approach makes it possible to study both the entire sample and a refined sample, filtered to focus on high-probability bad actors.

Although these longer-lived incidents may not be as time-sensitive as those monitored by live-feed, participants must still perform reverse DNS lookups and WHOIS queries on alleged perpetrator IP addresses and domain names as soon as possible after incidents are detected, not at the end of the study period.

A submission process will be designed to minimize participant effort while promoting consistent, accurate reporting. After a sufficiently large/broad set of third-party reports have been submitted, researchers will clean, code, and classify WHOIS data by Registrant Type as described in Section 2 for analysis as described in Section 4.

- **Phishing:** In proposal [8], the Anti Phishing Working Group (APWG) offered to supply a global list of phishing URLs, domains used to host them, and associated shutdown times. Due to the short duration of phishing sites, live-feed monitoring is preferable. However, analyzing this activity with both research methods might be useful to determine whether results differ significantly.

- **Cybersquatting:** Data on domains cited in alleged cybersquatting incidents might be gathered by organizations like the International Trademark Association (INTA). Approved dispute resolution service providers involved in ICANN's Uniform Domain-Name Dispute Resolution Policy (UDRP) are another possible source, although waiting until a dispute is filed to query WHOIS may be too much delay.

- **Intellectual property theft**: Data on domains cited in intellectual property theft complaints might be gathered by organizations like the UK Alliance Against IP Theft or the International Intellectual Property Rights (IPR) Advisory Program. However, data might be more readily available from groups that routinely record and investigate specific kinds of IP theft complaints, described below.

- **Media Piracy:** Data on domain names used by servers that illegally share copyrighted movies and music might be gathered by The International Federation of the Phonographic Industry (IFPI), the Motion Picture Association of America (MPAA), the Recording Industry Association of America (RIAA), and their international counterparts.

- **Software Piracy:** Data on domain names used by servers that illegally distribute copyrighted software might be gathered by major software vendors like Microsoft and Adobe or from an anti-piracy organization like the Business Software Alliance (BSA).

- **Trademark Infringement:** Data on domain names alleged to infringe upon registered trademarks might be gathered by an organization like the International Trademark Association (INTA) or commercial first-responders like Mark Monitor.

- **Counterfeit Merchandise:** Data on domains that send email advertising counterfeit merchandise and illegal pharmaceuticals might be gathered by an investigative agency like the US National Intellectual Property Rights Coordination Center Cyber Crimes Section (CCS). However, given that spam (one primary vector for online sale of counterfeit merchandise) can be studied more easily via live-feed, it might not be necessary to study this activity with method 2.

- **Money Laundering:** Data on domains that send recruiting email associated with fraudulent money laundering scams might be gathered by legitimate job recruitment websites like Monster and HotJobs or by an organization like BobBear that focuses specifically on tracking this type of illegal activity.

- **Advanced Fee Fraud:** Data on domains that send solicitation email associated with advanced fee fraud scams might be gathered by a tracking site like Artists Against 419 or bodies that handle Internet fraud complaints such as the FBI/NWCC Internet Crime Complaint Center (IC3) and its counterparts in other countries.

- **Identity Theft:** Data on domains that send bait email associated with online identity thefts might be gathered by the FBI/NWCC Internet Crime Complaint Center (IC3) or the US National Intellectual Property Rights Coordination Center Identity Fraud Initiative. However, major online identity theft vectors like phishing and malware can be studied more easily via live-feed monitoring; reliably correlating reported identity thefts to specific email messages and domains that caused them could be difficult.

- **Child Pornography:** Data on domain names of servers involved in online distribution of child pornography might be gathered by US National Intellectual Property Rights Coordination Center Cybercrimes Child Exploitation Section (CES) and Operation Predator. However, study [11] found it hard to obtain WHOIS data for child porn domains because, not only were sites taken down, but domain names were suspended.

- **Harassment or Stalking:** Input is requested on how to obtain a representative subsample of domain names that send online harassment and cyber-stalking email. Incidents are reported to local law enforcement agencies like FBI field offices. While HaltAbuse.org tracks statistics, based upon data supplied voluntarily by victims, many victims are reluctant to disclose these crimes. The highly personal nature of these activities could make it difficult to obtain a representative subsample.

- **Other Cybercrimes:** The FBI/NWCC Internet Crime Complaint Center (IC3) might also be able to supply data on perpetrator domains cited in complaints by victims of other cybercrimes, including online auction, investment fraud, and Internet extortion.

Because domain subsamples are likely to have some degree of cross-over, other readily-available online resources can be consulted to confirm and expand upon the kinds of illegal or harmful Internet communication associated with each domain. For example, in addition to RBLs, study [11] searched for domains using ReputationAuthority.org, Google Safe Browsing, McAfee SiteAdvisor, and Malware Domain List (either by searching a published list or by attempting to browse a website).

For each sampled domain, an **Apparent Registrant Type** must be assigned using the methodology defined by the WHOIS Accuracy Study [2], including confirmation of all domains potentially registered using Privacy/Proxy services. After this classification has been completed, the following input data will be available for each sampled domain:

Raw Data recorded by monitoring live-feed or reported by study participants

- Domain Name
- Registrant Name (may be a Privacy/Proxy service)
- Registrant Organization (may be a Privacy/Proxy service)
- Full WHOIS record for the domain
- Number of Illegal or Harmful Activity reported for this domain
- Kind(s) of Illegal or Harmful Activity reported for this domain
- Input Source(s) which supplied this domain name
- Incident Investigation Outcome (confirmed, refuted, in-progress/unknown)

Additional Data supplied by researchers

- Apparent Registrant Country Code/Name
- Apparent Registrant Type: missing/false, natural person, organization, multiple domain holder, or Privacy/Proxy service provider
- Additional Kind(s) of Illegal or Harmful Activity associated with this domain, as determined by searching RBLs and site reputation lists

4. Outputs

This study will quantify the frequency of Privacy/Proxy use among domains allegedly involved in illegal or harmful communication, broken down by kind of activity. To deliver these empirical results, this study will examine the WHOIS Registrant data associated with each sampled domain as follows.

- During classification, some domains will be found to have missing, patently false, or otherwise unusable WHOIS Registrant data, thereby impeding perpetrator identification. These domains represent another method of WHOIS abuse which should be measured and included in study findings, but do not constitute Privacy/Proxy abuse.

- During classification, some domains will be found to have WHOIS Registrant data that explicitly identifies and supplies direct contact information for a natural person, an organization (with or without a person's name), or a multiple domain holder. These Registrants may or may not actually be responsible for the reported illegal or harmful communication. For example, many domain names will be mapped to spambot-compromised residential broadband hosts or trojan-hacked websites operated by legitimate businesses. However, for the purposes of this study, the users of these domains shall be considered readily-identifiable and directly-contactable using Registrant data returned from a simple WHOIS query.

- The rest of the sample will consist of domains that, following classification, have WHOIS Registrant data that identifies an apparent Privacy/Proxy provider. For the purposes of this study, all such domains will be considered to have abused a Privacy/Proxy service for the purpose of obscuring perpetrator identification. To determine significance, this abuse rate shall be compared to the overall rate of Privacy/Proxy use measured by [3] (15-25%).

For each kind of activity studied, the following measurements will be derived from the entire subsample of alleged bad actors (including bots and other false positives):

- Percentage of entire sample that could not be analyzed, categorized by reason (e.g., false/missing WHOIS, recently modified WHOIS, suspended domain)
- Percentage of entire sample with Registrant NOT obscured via Privacy/Proxy, distributed by gTLD/country
- Percentage of entire sample apparently registered via Privacy service, distributed by gTLD/country
- Percentage of entire sample apparently registered via Proxy service, distributed by gTLD/country

For each kind of activity studied, similar measurements will also be derived from a refined subsample, filtered to reduce false positives and focus on confirmed bad actors:

- Percentage of refined sample that could not be analyzed, categorized by reason
- Percentage of refined sample with Registrant NOT obscured via Privacy/Proxy, distributed by gTLD/country
- Percentage of refined sample apparently registered via Privacy service, distributed by gTLD/country
- Percentage of refined sample apparently registered via Proxy service, distributed by gTLD/country

Finally, these results will be aggregated and used to answer the following questions:

- Are Privacy services abused more/less often by bad actors (alleged or confirmed)?
- Are Proxy services abused more/less often by bad actors (alleged or confirmed)?
- Which illegal/harmful activities are most likely to abuse Privacy/Proxy services?
- Which illegal/harmful activities are least likely to abuse Privacy/Proxy services?
- Were there any kinds of illegal/harmful Internet communication for which Privacy/Proxy abuse could not be studied in a reliable way and why?

5. References

[1] Working Definitions for Key Terms that May be Used in Future WHOIS Studies, GNSO Drafting Team, 18 February 2009

[2] Proposed Design for a Study of the Accuracy of Whois Registrant Contact Information (6558,6636), NORC, June 3, 2009

[3] ICANN's Study on the Prevalence of Domain Names Registered using a Privacy or Proxy Service among the top 5 gTLDs, ICANN, September 28, 2009

[4] Registrar Accreditation Agreement (RAA), ICANN, 21 May 2009

[5] Terms of Reference for WHOIS Misuse Studies, ICANN, September 2009

[6] Terms of Reference for WHOIS Registrant Identification Studies, ICANN, Oct 2009

[7] Terms of Reference for WHOIS Privacy/Proxy Reveal Studies, ICANN, In Progress

[8] Study Suggestion Number 13b/c, Measure growth of proxy/privacy services vis-à-vis all registrations, Laura Mather

[9] Study Suggestion Number Study 17, Identify why proxy/privacy service users use those services, Claudio DiGangi

[10] GAC Data Set 11, What is the percentage of domain names registered using proxy or privacy services that have been associated with fraud or other illegal activity, GAC Recommendations for WHOIS Studies, 16 April 2008

[11] Private Domain Registrations: Examining the relationship between private domain registrations and malicious domains at 3FN, Piscitello, October 2009