Statistical Analysis of DNS Abuse in gTLDs (SADAG)
Background, Methodology, and Planned Research
SIDN and Delft U. of Technology | ICANN 58 | 10 March 2017
Agenda

⊙ Introduction from ICANN: Background of Study

⊙ Presentation on Methodology and Planned Research from SIDN and Delft University of Technology (TU-Delft)

⊙ Q & A
## 2009

**Mitigating Malicious Conduct: New gTLD Program Explanatory Memorandum**

<table>
<thead>
<tr>
<th>Question</th>
<th>Recommendation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) How do we ensure that bad actors do not run registries?</td>
<td>1. Vet registry operators</td>
</tr>
<tr>
<td>2) How do we ensure integrity and utility of registry information?</td>
<td>2. Require DNSSEC Deployment</td>
</tr>
<tr>
<td></td>
<td>3. Prohibit “wildcarding”</td>
</tr>
<tr>
<td></td>
<td>4. Encourage removal of “orphan glue” records</td>
</tr>
<tr>
<td>3) How do we ensure more focused efforts on combating identified abuse?</td>
<td>5. Require “Thick” WHOIS records</td>
</tr>
<tr>
<td></td>
<td>6. Centralize Zone File access</td>
</tr>
<tr>
<td></td>
<td>7. Document registry- and registrar-level abuse contacts and policies</td>
</tr>
<tr>
<td></td>
<td>8. Provide an expedited registry security request process</td>
</tr>
<tr>
<td>4) How do we provide an enhanced control framework for TLDs with intrinsic potential for malicious conduct?</td>
<td>9. Create a draft framework for a high security zone verification program</td>
</tr>
</tbody>
</table>
Study Background

2016

- New gTLD Program Safeguards Against DNS Abuse: Revised Report
- Research aid to Competition, Consumer Trust, and Choice Review Team
- How to measure effectiveness of safeguards?

Base Research Model

Explanatory Variable: DNS Expansion

Potential proxy metrics:
- Number of domain names
  - Legacy TLDs
  - New TLDs
  - Entire DNS

Response Variable: DNS Abuse Rate

Potential proxy metrics:
- Spam rate
- Phishing rate
- Malware rate
- Others as relevant to the “effectiveness” objectives of the safeguards

Intervening Variable(s)
Safeguards to Mitigate DNS Abuse

What about…
- Pricing?
- Operational policies and/or practices?
- Systemic policies and/or practices?
- Cybercriminal preferences and practices?
Competition, Consumer Choice, and Trust Review Team

Mandated by AoC to examine “effectiveness of…safeguards put in place to mitigate issues involved in…the expansion [of the top-level domain space]”

Required comprehensive descriptive statistics as baseline measure of abuse rates in new compared to legacy gTLDs

CCTRT recommends ongoing measurement of abuse to answer fundamental question:

What explains the variation in abuse rates in TLDs?

- RFP issued August 2016
- SIDN contracted November 2016
- Research began December 2016
- Final report expected June 2017

Big Project!
Tight Timeframe!
Need Data!
Statistical Analysis of DNS Abuse in gTLDs (SADAG)
Methodology and Planned Research
Maarten Wullink – SIDN
Maciej Korczyński – Delft U. of Technology
Project

Statistical Analysis of DNS Abuse in gTLDs (SADAG)

Consortium: SIDN and TU Delft

Requested by: Competition, Consumer Choice, and Trust Review Team
Goal

– Comprehensive statistical comparison of rates of DNS abuse in new and legacy gTLDs
  ▪ Spam
  ▪ Phishing
  ▪ Malware
  ▪ Botnet Command-and-Control

– Statistical analysis of potential relationship with abuse drivers
  • DNSSEC
  • Other drivers as identified by future Review Teams
Motivation

– New Generic Top-Level Domain (gTLD) Program enabled hundreds of new generic top-level domains

– Safeguards built into the Program intended to mitigate rates of abusive, malicious, and criminal activity in these new gTLDs
Current data providers (1)

Domain Blacklists

- Anti Phishing Working Group
  • Phishing URLs

- StopBadware
  • Malware URLs

- Secure Domain Foundation
  • Malware URLs (Command & Control, EXE, Compromised)
  • Phishing URLs
  • Highly suspect domains
  • Bad Faith domains
Current data providers (2)

WHOIS data

- Whois XML API
  - All new gTLDs
  - Subset of legacy gTLDs

Domain data

- Zone files
  - Per gTLD
  - Per day
  - 3 year period
gTLD groups

Legacy gTLDs
- E.g. .com, .org, .net, asia, .biz etc.

New gTLDs
- Part of the New gTLD program
- E.g. amsterdam, .xyz

<table>
<thead>
<tr>
<th>Study component</th>
<th># Legacy gTLDs</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLD level aggregation</td>
<td>17</td>
<td>Zone files</td>
</tr>
<tr>
<td>Maliciously registered vs. compromised domains</td>
<td>9</td>
<td>WHOIS data</td>
</tr>
<tr>
<td>Registrar aggregation</td>
<td>9</td>
<td>WHOIS data</td>
</tr>
</tbody>
</table>
Data limitations

WHOIS data

- Collection method
  • No continuous scanning
  • Might be missing short-lived domains
More Data Requested!

- Abuse feeds
  - Phishing
  - Malware
  - Botnet C&C
  - Spam

- Uptimes
Security metrics

- Concentration of malicious content:
  - Number of unique domains
    - E.g. malicious.com
Security metrics

– Concentration of malicious content:
  • Number of unique domains
    – E.g. malicious.com
  • Number of FQDNs
    – E.g. 123.malicious.com, 456.malicious.com, 789.malicious.com, (…)

Statistical Analysis of DNS Abuse in gTLDs
Security metrics

– Concentration of malicious content:
  • Number of unique domains
    – E.g. malicious.com
  • Number of FQDNs
    – E.g. 123.malicious.com, 456.malicious.com, 789.malicious.com, (…)
  • Number of URLs
    – E.g. malicious.com/wp-content/file.php, malicious.com/wp-content/gate.php, (…)
<table>
<thead>
<tr>
<th>STOP BADWARE (SITES)</th>
<th>F.I.R.E. (COMPOSITE)</th>
<th>PHISHTANK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planet.com (AS21844)</td>
<td>ThePlanet.com (AS21844)</td>
<td>NJ INTL INTERNET EXCHANGE (AS16812)</td>
</tr>
<tr>
<td>JANET BACKBONE (AS14035)</td>
<td>PAH Inc GoDaddy.com (AS26496)</td>
<td>MetroRED Telecom Services (AS13591)</td>
</tr>
<tr>
<td>Inc GoDaddy.com (AS26496)</td>
<td>OVH - OVH (AS16276)</td>
<td>RAPIDSWITCH-AS (AS29131)</td>
</tr>
<tr>
<td>Google Inc. (AS6151)</td>
<td>BLUEHOST-AS (AS11798)</td>
<td>CENTROHOST-AS (AS41126)</td>
</tr>
<tr>
<td>Layer Technologies (AS36351)</td>
<td>IPNAP- GigeNET (AS23522)</td>
<td>ThePlanet.com (AS21844)</td>
</tr>
<tr>
<td>Sprint Co/PSI (AS174)</td>
<td>EcomD-Coloquest/GigeNet (AS32181)</td>
<td>iWeb Technologies Inc. (AS32613)</td>
</tr>
<tr>
<td>NET Beijing (AS17431)</td>
<td>GNAXNET - Global Net Access (AS3595)</td>
<td>Softlayer Technologies (AS36351)</td>
</tr>
<tr>
<td>African Internet Svcs (AS6130)</td>
<td>Bizland-SD - Endurance Intl (AS29873)</td>
<td>OVH - OVH (AS16276)</td>
</tr>
<tr>
<td>ARBOR TOP ASN THREATS</td>
<td>EMERGING THREATS COMPROMISED IPS</td>
<td>Limestone Networks Inc (AS46475)</td>
</tr>
<tr>
<td>CHINA TELECOM (AS4134)</td>
<td>Emerging Threats RBN</td>
<td>SOVAM-AS Golden Telecom (AS3216)</td>
</tr>
</tbody>
</table>

Security metrics

Size matters!
Size estimates

- Size of a TLD can be used as an explanatory factor for the concentrations of abused domains

- Size of a TLD could be interpreted as the “attack surface” size for cybercriminals.

- Number of 2\textsuperscript{nd}-level domains registered in each gTLD (zone files)

- Limitation: There is a large portion of domains in new gTLDs with NS records that do not resolve yet
  
  - Solution: active measurement to determine domains in use per gTLD
Size estimates

- Number of 2nd-level domains registered in each registrar (WHOIS data)

- Limitation: single entity can have multiple different names, e.g., we found a registrar using 52 distinct name variations
  - Solution: an additional entity resolution step to try to group together the different names of a single registrar (58% reduction)

- Limitation: missing WHOIS data
Compromised versus maliciously registered domains

- Definitions:
  - Maliciously registered domain – domain registered by a miscreant for malicious purposes
  - Compromised domain – domain registered by a legitimate user and hacked by a miscreant
  - Third party domains – domains of legitimate services that tend to be misused by miscreants (e.g. file sharing services, blog post services, URL shortening services)

- For compromised domains, the TLD size could be interpreted as the “attack surface” size for cybercriminals.

- For malicious registrations, the TLD size could serve as a proxy for the “popularity” of the TLD. What makes it popular?
Distinguishing between compromised and maliciously registered domains

- Distinguishing between compromised and maliciously registered domains is critical because they require different mitigation actions by different intermediaries.

- Assumption: maliciously registered domains are involved in a criminal activity within a short time after the registration.

- Limitation: (lack of) WHOIS data, maliciously registered domains involved in a criminal activity within a longer time after the registration, or delayed blacklisting.
  
  • Solution: more advanced machine learning approach (requires more “features” and the “ground truth” data).
Future work

– Incorporate more blacklist feeds

– Analyze abuse per:
  • Reseller
  • Privacy / proxy service (if data available)
  • Geographic region

– Analysis of the time-to-live of domain names
  • Requires uptime data

– Inferential analysis of potential relationship with abuse drivers
Schedule

- Final report available early June 2017
Questions?