DNS and Evidence-Based Security

WIE-KISMET

December 9, 2019

Geoffrey M. Voelker

University of California, San Diego

UCSDCSE
Computer Science and Engineering

Evidence-Based Security

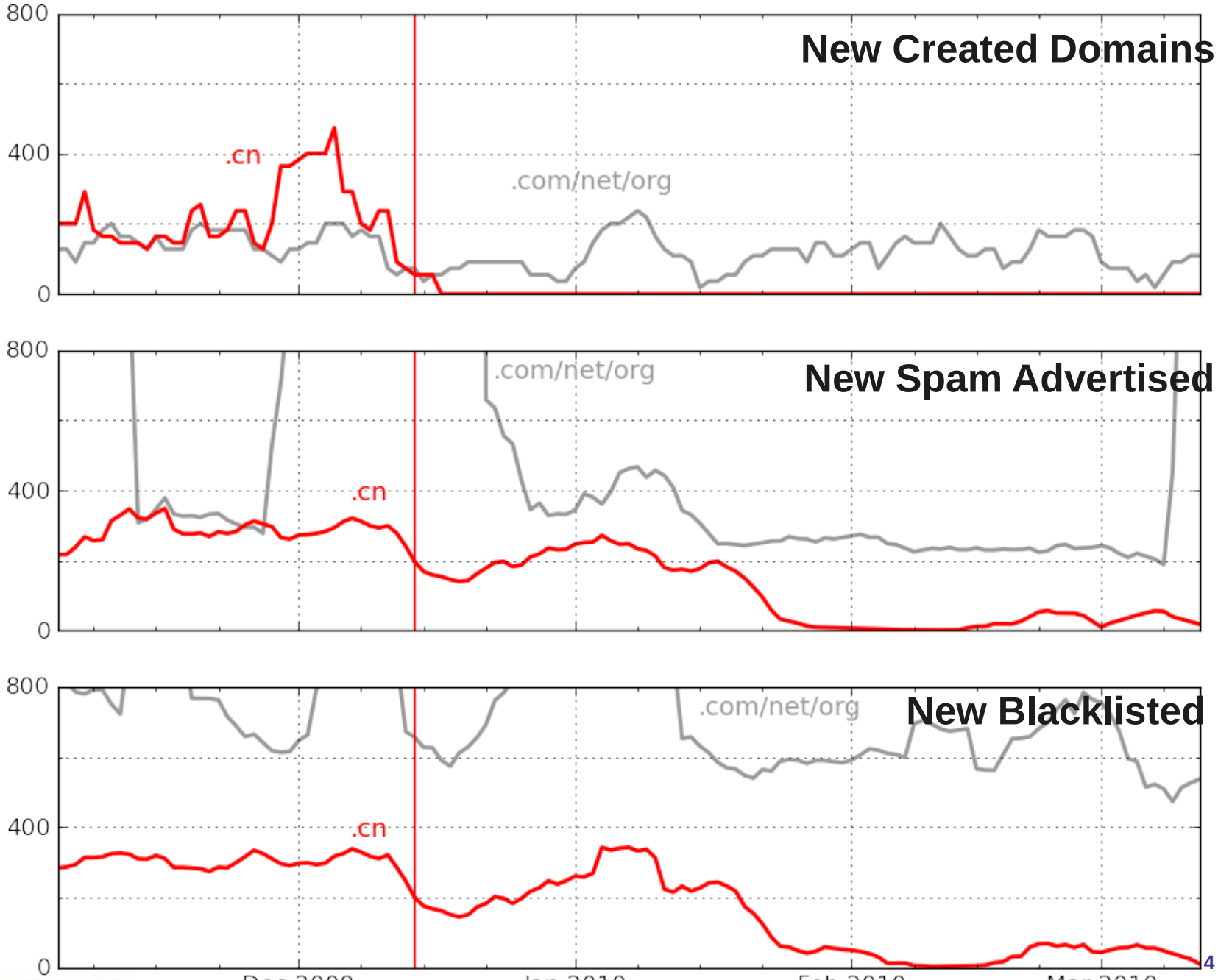
- Our work in DNS and related areas has been motivated by long-term cybersecurity projects
  - Wide variety of security projects over time
  - DNS often plays a role since it is a fundamental resource
- Our approach has been heavily measurement-based
  - Effective intervention requires reasoning about motivations, incentives, requirements, communities

Impact of Domain Registration Policy Changes

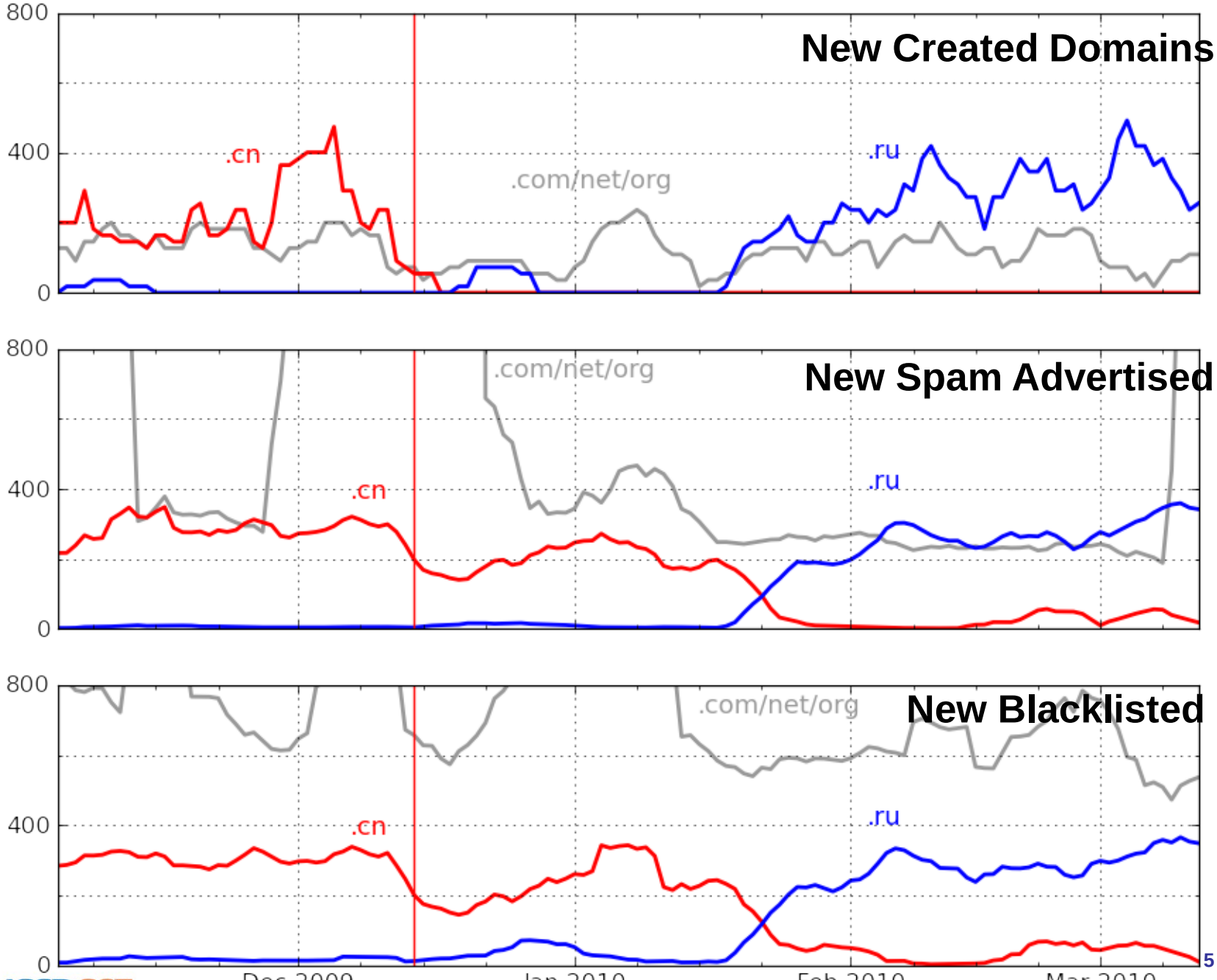
- Dec 2009: CCNIC policy changes induces 70x change in price of .cn domains

- Effectively, a global sweeping change by a registrar
- How did that influence spammers?

Liu, Levchenko, Félegyházi, Kreibich, Maier, Voelker, Savage, On the Effects of Registrar-level Intervention, LEET 2011



UCSDCSE
Computer Science and Engineering



UCSDCSE
Computer Science and Engineering

Impact of New TLDs

- Explore impact of new TLDs on DNS
- Do new TLDs serve their purpose (“meet unmet needs”)?
- Approach
  - Examine one new TLD in detail
  - Expand to all new TLDs (circa 2014)

The .xxx TLD

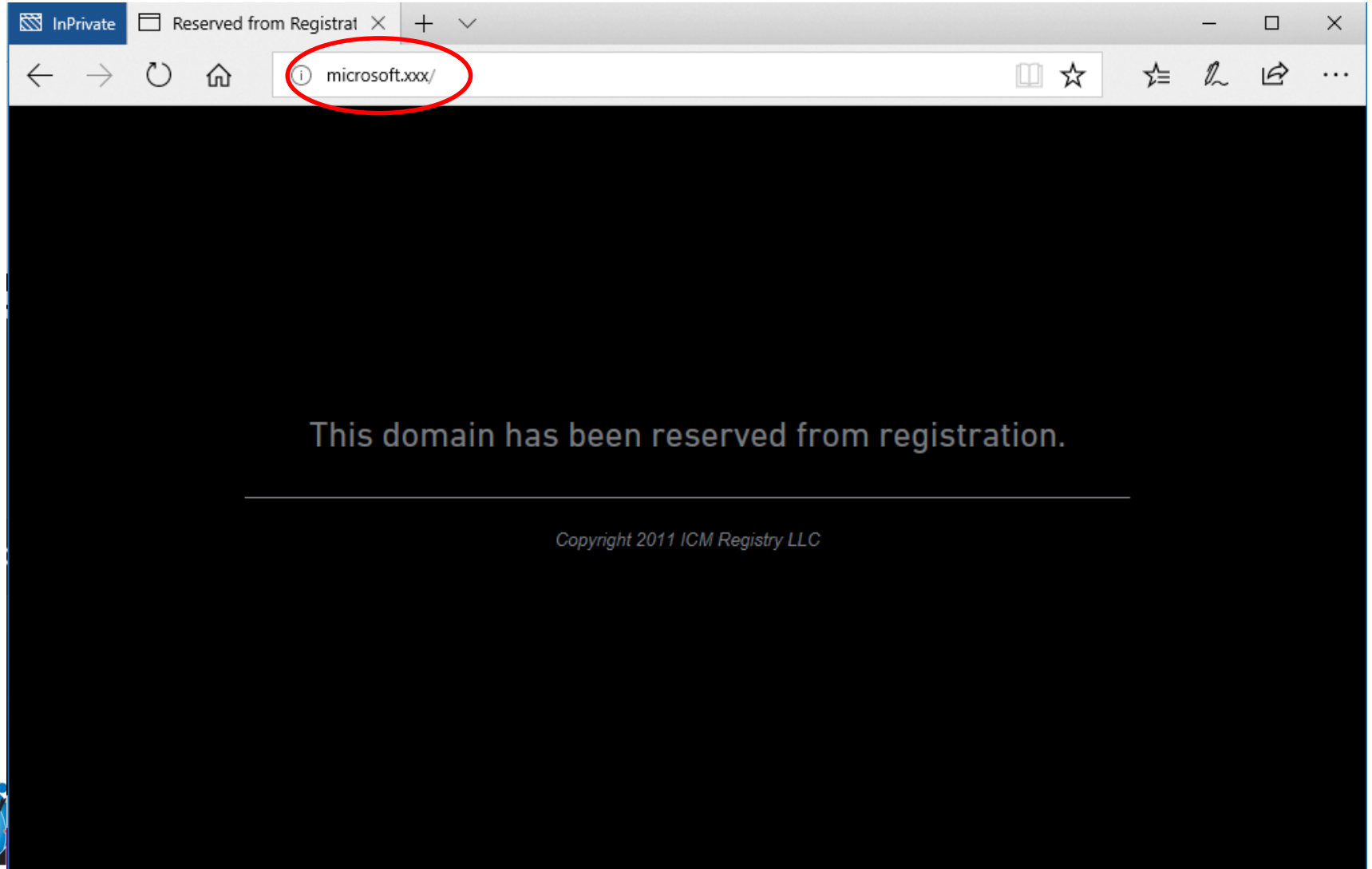
- Unusual TLD with storied history
- Specialized TLD intended for adult content
  - First proposed in 2000 by ICM Registry
  - Debated for 10 years
  - "…community will consist of the responsible global online adult-entertainment community"
- Criticisms from many parties
  - Trademark holders
  - Adult entertainment industry (Free Speech Coalition)

Halvorson, Levchenko, Savage, Voelker, XXXtortion? Inferring Registration Intent in the .XXX TLD, WWW 2014

Content Categorization

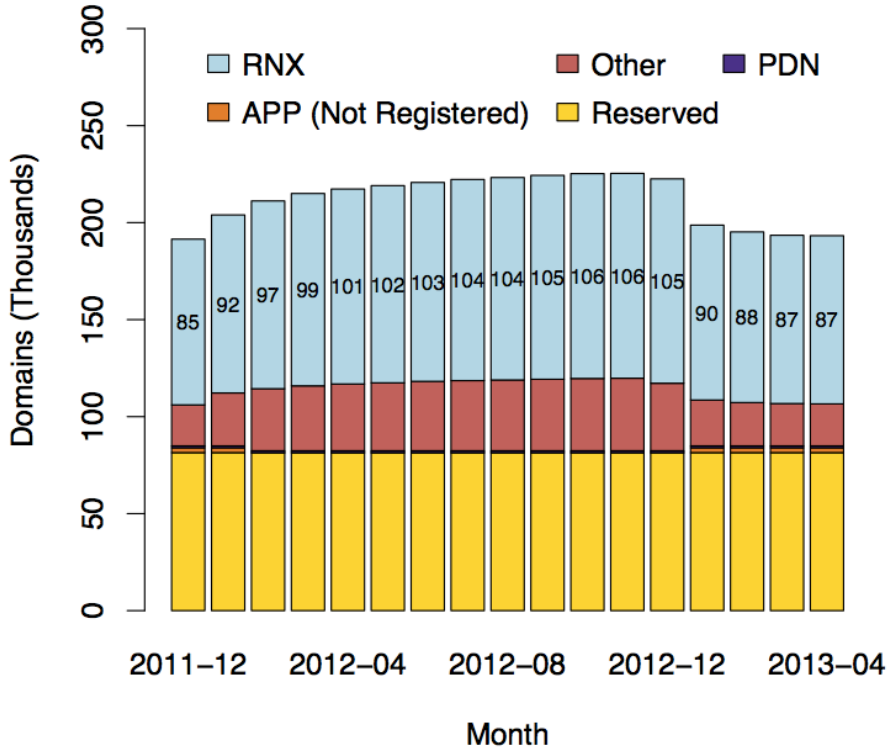
- Classified all .xxx domains by type of content served
  - 193,363 domains in April 2013
- Web content
  - Crawled all domains in zone file
  - January 10, 2013 and April 12, 2013
  - Clustered using text shingling
  - Generate labels using top clusters
- WHOIS records
  - For identifying registered non-resolving

Reserved Domains



Registered Non-Resolving

- Registered but not in zone
  `% dig ucsd.xxx → NXDOMAIN`
- GoDaddy: “this is how to defend”
- Use ICANN reports
  - No exhaustive list
  - Can infer numbers
- Intent: Defensive



Summary

- Does .xxx meet unmet needs?

→ Absolutely not

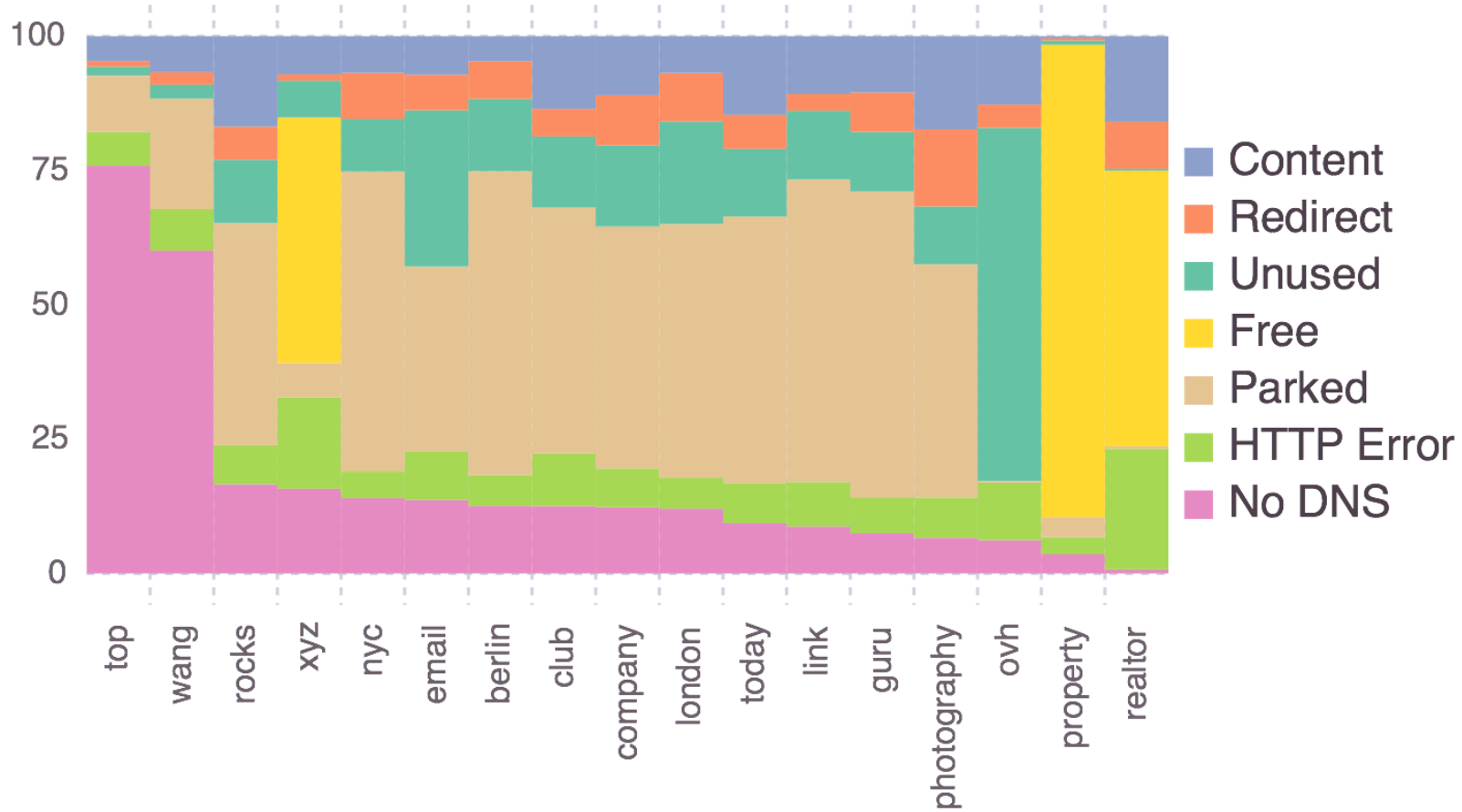
- Little benefit to intended demographic
  - Whatever adult content is out there, it's not in .xxx
- Huge cost to everyone else
  - Defensive registrations 93% of ongoing revenue
  - To protect yourself, you have to register to prevent someone else from registering it for you

New gTLDs

- Comprehensively identify all domains in new TLDs
  - New TLDs up to 2015
  - Register for zone file access at ICANN
  - Download over 500 zone files daily
- DNS + Web crawl for content
  - Every domain in a new TLD
  - Millions from old TLDs (for reference)
  - Web: 150GB visit, 1.5TB screenshots
- Cluster + label downloaded content
  - Bag of words, k-means, active learning

Halvorson, Der, Foster, Savage, Saul, Voelker, From .academy to .zone: An Analysis of the New TLD Land Rush, IMC 2015

Content in Top TLDs



Registration Intent

Registration Intent	Result	
Primary	378,401	14.9%
Defensive	1,005,109	39.5%
Speculative	1,161,892	45.6%

Primary registrations the lowest category

Registrar-level Attacks

- Recently we have been interested in registrar attacks
  - Registrar compromise, registrar account compromise, etc.
- Attackers gain substantial leverage
  - Shadow subdomains, DNS hijacking, etc.
  - Motivated by attacks such as the 2014 Snecma.fr attack
  - Particularly problematic since changes come from “owner”
- Have been focusing on nameservers in particular
  - Valuable targets, particularly useful for hijacking

Nameserver Abuse

- Initially focused on suspicious nameserver activity
  - Active crawls and passive zone files
- But unusual behaviors can have benign explanations
  - New NS added for 1-2 days that maps to an unusual /24?
  - Sometimes highly suspicious…sometimes benign
- Have been systematically categorizing nameserver dynamics to establish a “baseline”
  - Consistency
    - Misconfigurations, incomplete data, routing issues, etc.
  - Diversity
    - Topological concentration of NS’s and domains that use them
  - Dynamics
  - Joint with University of Twente, CAIDA, Ian Foster

Threat Intel

- Threat Intelligence (TI) feeds distribute “indicators of compromise” for input into defenses
  - IP addresses, file hashes, domain names, URLs
  - Appearing on a feed indicates something “bad”
- Using feeds now a standard operational practice
  - Many feed sources, both public and commercial
- How can a user evaluate the quality and utility of threat intelligence feeds?
  - How do you choose which feed to use, or how many?
  - How useful are they? (How do you define useful?)

Li, Dunn, Pearce, McCoy, Voelker, Savage, Levchenko, Reading the Tea Leaves: A Comparative Analysis of Threat Intelligence, USENIX Security 2019

Threat Intel Evaluation

- Define six metrics for evaluation
  - Volume, differential contribution, exclusive contribution, latency, accuracy, coverage
- Define methods for calculating metrics across feeds
  - Account for variations (e.g., snapshot vs event)
- Examine 47 IP feeds and 8 malware hash feeds
  - Dec 2017 – July 2018
  - Commercial and public feeds
  - Categorized into six types: scan, brute force, malware, botnet, exploit, spam

Threat Intel Results

- Significant issues across the metrics
  - Coverage is poor when compared to ground truth data
    - Scan feeds all combined only account for 2% of telescope scans
  - Accuracy issues can lead to false positives
    - Non-trivial amount of unroutable, top Alexa, CDN IPs
  - Most IP indicators are singletons (very low intersection)
  - Little evidence that larger feeds contain better data
- Challenges
  - Providers do not explain how data is collected and labelled
    - Left to users to decide how to interpret
  - Little insight into operational uses of feeds