

Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services

Pelayo Vallina, Victor Le Pochat, Álvaro Feal,
Marius Paraschiv, Julien Gamba, Tim Burke,
Oliver Hohlfeld, Juan Tapiador, Narseo Vallina-Rodriguez





[NBA Draft](#)
[Wimbledon](#)

ONSALE

FIVE \$1,000
Winners Click Now!

[Holyfield](#)
[vs. Tyson](#)

 [options](#)

[Yellow Pages](#) - [People Search](#) - [Maps](#) - [Classifieds](#) - [News](#) - [Stock Quotes](#) - [Sports Scores](#)

- ◆ [Arts and Humanities](#)
[Architecture](#), [Photography](#), [Literature](#)...
- ◆ [Business and Economy \[Xtra!\]](#)
[Companies](#), [Investing](#), [Employment](#)...
- ◆ [Computers and Internet \[Xtra!\]](#)
[Internet](#), [WWW](#), [Software](#), [Multimedia](#)...
- ◆ [Education](#)
[Universities](#), [K-12](#), [College Entrance](#)...
- ◆ [Entertainment \[Xtra!\]](#)
[Cool Links](#), [Movies](#), [Music](#), [Humor](#)...
- ◆ [News and Media \[Xtra!\]](#)
[Current Events](#), [Magazines](#), [TV](#), [Newspapers](#)...
- ◆ [Recreation and Sports \[Xtra!\]](#)
[Sports](#), [Games](#), [Travel](#), [Autos](#), [Outdoors](#)...
- ◆ [Reference](#)
[Libraries](#), [Dictionaries](#), [Phone Numbers](#)...
- ◆ [Regional](#)
[Countries](#), [Regions](#), [U.S. States](#)...
- ◆ [Science](#)
[CS](#), [Biology](#), [Astronomy](#), [Engineering](#)...

web directories
(manually edited)

Web Content Filtering

Choose your filtering level

- High** Protects against all adult-related sites, illegal activity, social networking sites, video sharing sites, and general time-wasters.
26 categories in this group - [View](#) - [Customize](#)
- Moderate** Protects against all adult-related sites and illegal activity.
13 categories in this group - [View](#) - [Customize](#)
- Low** Protects against pornography.
4 categories in this group - [View](#) - [Customize](#)
- None** Nothing blocked.
- Custom** Choose the categories you want to block.

- | | | |
|---|---|---|
| <input type="checkbox"/> Academic Fraud | <input type="checkbox"/> Adult Themes | <input type="checkbox"/> Adware |
| <input type="checkbox"/> Alcohol | <input type="checkbox"/> Anime/Manga/Webcomic | <input type="checkbox"/> Auctions |
| <input type="checkbox"/> Automotive | <input type="checkbox"/> Blogs | <input type="checkbox"/> Business Services |
| <input type="checkbox"/> Chat | <input type="checkbox"/> Classifieds | <input type="checkbox"/> Dating |
| <input type="checkbox"/> Drugs | <input type="checkbox"/> Ecommerce/Shopping | <input type="checkbox"/> Educational Institutions |
| <input type="checkbox"/> File Storage | <input type="checkbox"/> Financial Institutions | <input type="checkbox"/> Forums/Message boards |
| <input type="checkbox"/> Gambling | <input type="checkbox"/> Games | <input type="checkbox"/> German Youth Protection |

classification engines
(automated)

Why does the quality of these services matter?

- › *End users*: incorrect categories affect reliability
 - ›› over/underblocking in content filtering
- › *Academia*: domain sample or results depend on them
 - ›› 2019 top conferences: 24 papers
 - ›› lack of trust → resort to manual classification

Services are opaque on how they operate

The Forcepoint Master Database contains the industry's most accurate, current and comprehensive classification of URLs. We use proprietary classification software and

Webshrinker uses advanced Machine Learning algorithms.

Collect the best websites for any topic!

Validation? Training set? Comprehensiveness?

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Inputs

Outputs

Purpose

Updates

Access



Inputs

Outputs

Purpose

Updates

Access



OpenDNS

Aggregator



Inputs

Outputs

Purpose

Updates

Access



Inputs

Outputs

Purpose

Updates

Access

Content filtering

Threat assessment



FortiGuard Labs
Global threat research and response

Bitdefender®



VIRUSTOTAL



McAfee™



Dr.WEB®



Alexa



Symantec



TREND MICRO™



dmoz



Forcepoint websense



webshrinker



OpenDNS



Curlie 🌰

Marketing

Discovery

Inputs

Outputs

Purpose

Updates

Access

(Mostly) automated



OpenDNS



Manual

Inputs

Outputs

Purpose

Updates

Access



Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

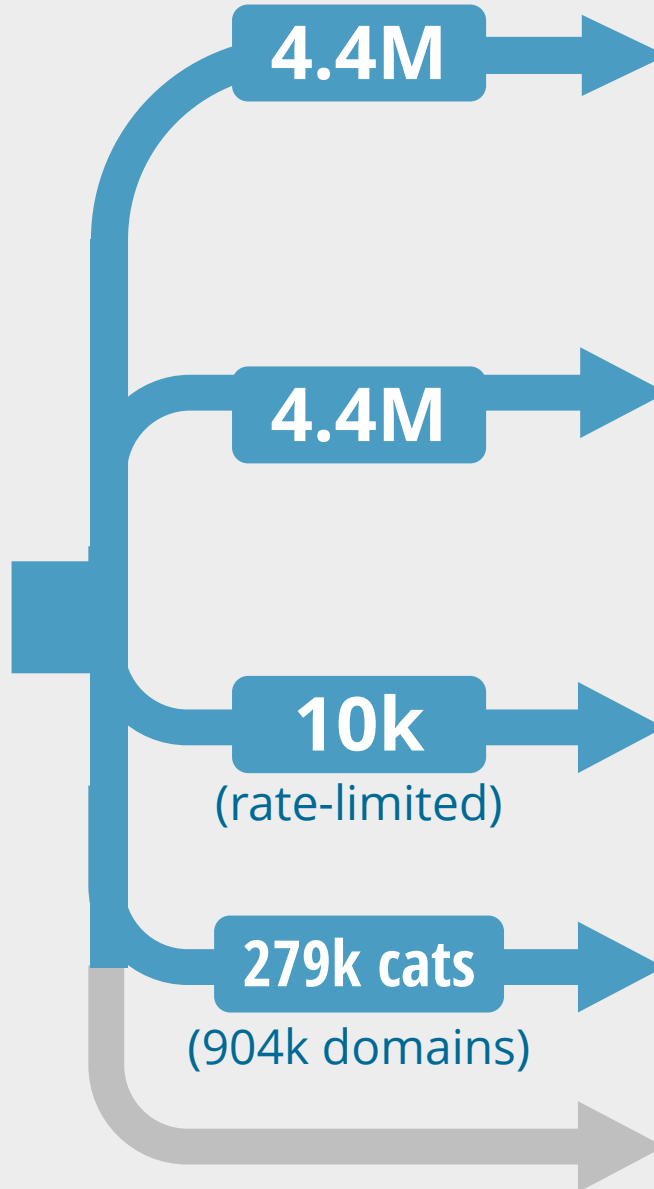
**Discussion
&
Conclusion**

Label gathering

 **Alexa**
Top 1M domains
Sept 1-30
2019





Aggregate
using
Tranco
↓
4.4M
domains






FortiGuard Labs
Global threat research and response


McAfee™ OpenDNS


VIRUSTOTAL



 **Alexa** Bitdefender  **Dr.WEB®**

Forcepoint websense  **TREND MICRO™**

 **Symantec**  webshrinker

 **TREND MICRO™** *direct*

 **Alexa** *direct*

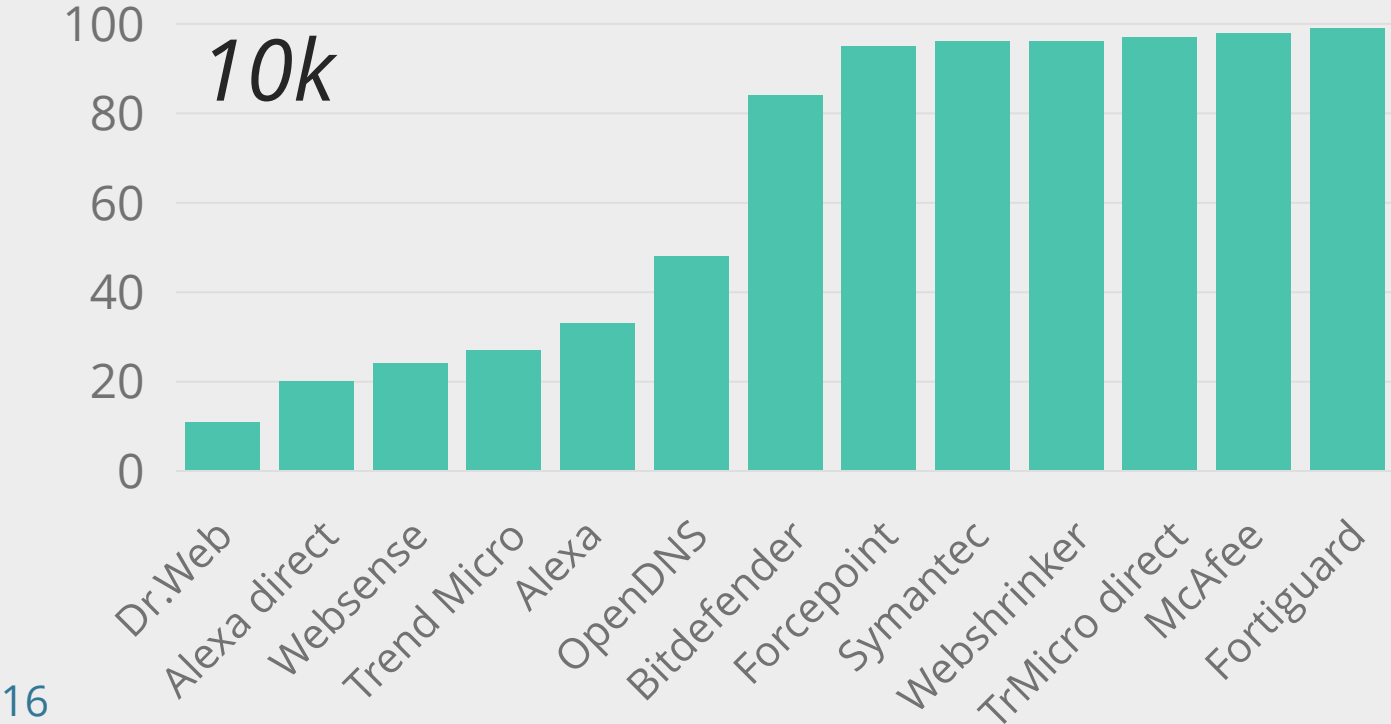
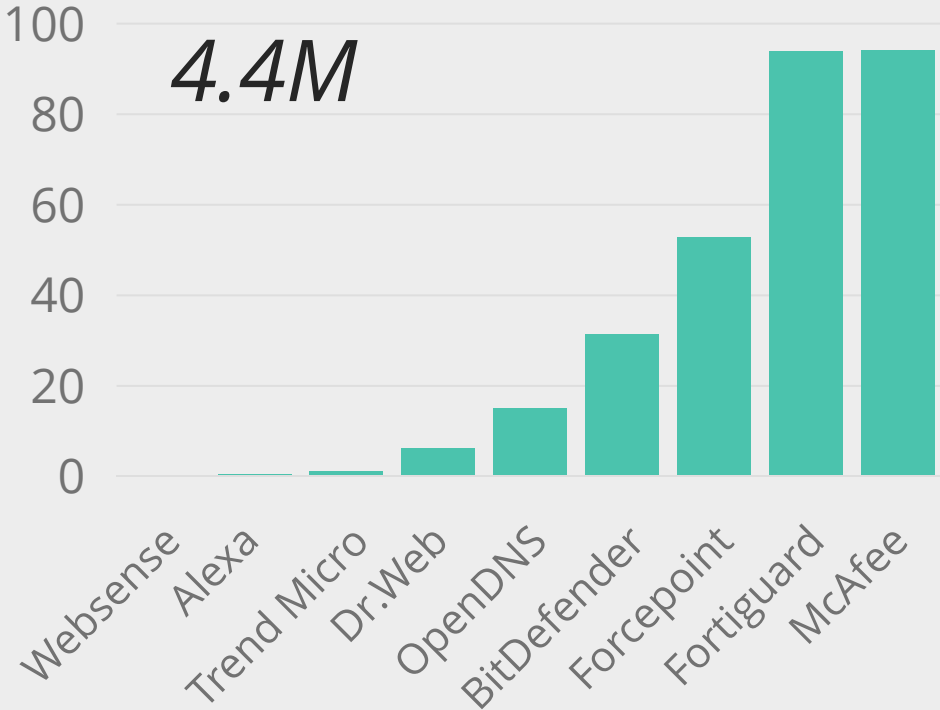
Curlie  

Service choice affects which domains are labeled

› Coverage

- ›› ranges from <1% to 94%
- ›› is better for automated classification services

Updates



Service choice affects which domains are labeled

› Coverage

- ›› ranges from <1% to 94%
- ›› is better for automated classification services

› Popular domains have better coverage

› Subdomain coverage ranges from <1% to 99%

› Inconsistent when directly sourced or through VirusTotal

Updates

Updates

Inputs

Access

Service choice affects the taxonomy granularity

- › Security/content filtering: **fewer** categories
 - › As low as 12
 - › Easier setup
- › Marketing: **more** categories
 - › Up to 7.5k
 - › Fine-grained targeting

Purpose

Service choice affects label interpretation

- › **Inconsistencies** between documented and observed labels
- › Multiple labels are **uncommon**
- › Subdomains **inherit** labels from parent
- › 3 out of 9 services **updated** labels
 - ›› Mostly for maliciousness

Access

Outputs

Inputs

Updates

Service choice affects label distribution

- › **Disagreement**

 - on distribution of labels over domains

 - ›› As measured through *mutual information*

- › **Uneven** distribution of labels over domains

 - ›› As measured through *label frequency*

Purpose

Updates

Purpose

- mcf_sports
- mcf_general news
- mcf_provocative attire
- mcf_for kids
- mcf_streaming media
- mcf_resource sharing
- mcf_health
- mcf_anonymizers
- mcf_politics/opinion
- mcf_blogs/wiki
- mcf_media sharing
- mcf_entertainment
- mcf_portal sites
- mcf_games
- mcf_motor vehicles
- mcf_job search
- mcf_public information
- mcf_illegal software
- mcf_marketing/merchandising
- mcf_technical information
- mcf_stock trading
- mcf_web mail
- mcf_fashion/beauty
- mcf_online shopping
- mcf_finance/banking
- mcf_technical/business forums
- mcf_internet services
- mcf_business
- mcf_pornography
- mcf_software/hardware
- mcf_parked domain
- mcf_shareware/freeware
- mcf_instant messaging
- mcf_chat
- mcf_real estate
- mcf_media downloads
- mcf_travel
- mcf_search engines
- mcf_text translators
- mcf_internet radio/iv
- mcf_education/reference
- mcf_personal network storage
- mcf_government/military

McAfee

- odns_sports
- odns_politics
- odns_news/media
- odns_health and fitness
- odns_proxy/anonymizer
- odns_radio
- odns_movies
- odns_music
- odns_television
- odns_automotive
- odns_blogs
- odns_portals
- odns_games
- odns_jobs/employment
- odns_photo sharing
- odns_be the first to tag this domain
- odns_p2p/file sharing
- odns_forums/message boards
- odns_webmail
- odns_ecommerce/shopping
- odns_lingerie/bikini
- odns_advertising
- odns_anime/manga/webcomic
- odns_nudity
- odns_sexuality
- odns_financial institutions
- odns_business services
- odns_instant messaging
- odns_pornography
- odns_software/hardware
- odns_chat
- odns_adware
- odns_podcasts
- odns_software/technology
- odns_visual search engines
- odns_research/reference
- odns_video sharing
- odns_classifieds
- odns_search engines
- odns_travel
- odns_government
- odns_educational institutions
- odns_file storage

OpenDNS

- bdf_sports
- bdf_news
- bdf_portals
- bdf_radiomusic
- bdf_entertainment
- bdf_health
- bdf_marketing
- bdf_games
- bdf_parked
- bdf_misc
- bdf_blogs
- bdf_business
- bdf_drugs
- bdf_online shop
- bdf_webmail
- bdf_porn
- bdf_computers and software
- bdf_search engines
- bdf_hosting
- bdf_social networks
- bdf_financial
- bdf_education
- bdf_hobbies
- bdf_file sharing
- bdf_government

BitDefender

- fp_news and media
- fp_entertainment
- fp_sports
- fp_proxy avoidance
- fp_society and lifestyles
- fp_web analytics
- fp_adult content
- fp_instant messaging
- fp_search engines and portals
- fp_collaboration - office
- fp_text and media messaging
- fp_health
- fp_streaming media
- fp_job search
- fp_web and email marketing
- fp_vehicles
- fp_games
- fp_shopping
- fp_business and economy
- fp_uncategorized
- fp_general email
- fp_web collaboration
- fp_blogs and personal sites
- fp_message boards and forums
- fp_information technology
- fp_sex
- fp_financial data and services
- fp_educational materials
- fp_hosted business applications
- fp_hacking
- fp_media file download
- fp_real estate
- fp_educational institutions
- fp_travel
- fp_personal network storage and backup
- fp_government
- fp_organizational email

ForcePoint

- vt_news and media
- vt_twiste_electronique
- vt_uz
- vt_weather
- vt_zhifeng
- vt_zhifeng Nachrichten
- vt_pinnacolo
- vt_salle_pict
- vt_messagers
- vt_jf99a
- vt_news media
- vt_sports
- vt_sports_recreation hobbies
- vt_reviews
- vt_zpravodajstvi
- vt_arts_entertainment
- vt_zeitschriften und online-magazine
- vt_walt_disney_pictures
- vt_social networks
- vt_online_communities
- vt_radiomusic
- vt_letras
- vt_news
- vt_news media entertainment
- vt_entertainment
- vt_portals
- vt_web advertisement
- vt_proxy avoidance
- vt_search engines and portals
- vt_web analytics
- vt_search engines portals
- vt_portals
- vt_instant messaging
- vt_provedores de acesso
- vt_wifi_aveiroptics
- vt_parked
- vt_text and media messaging
- vt_marketing
- vt_society and lifestyles
- vt_collaboration - office
- vt_新聞とIT
- vt_streaming media
- vt_health ストメディア
- vt_job search
- vt_games
- vt_job search careers
- vt_employment
- vt_misc
- vt_search engines
- vt_vehicles
- vt_web and email marketing
- vt_tickets
- vt_business economy
- vt_automotive
- vt_shopping
- vt_fertigbau
- vt_onlineshop
- vt_v
- vt_business and economy
- vt_women's
- vt_jobs
- vt_information services
- vt_news_and_media
- vt_email
- vt_business
- vt_computers internet
- vt_general email
- vt_email
- vt_webmail
- vt_adult content
- vt_message boards and forums
- vt_uncategorized
- vt_advertisements
- vt_web collaboration
- vt_hosting
- vt_streaming media mp3
- vt_blogs
- vt_information technology
- vt_social networks
- vt_computers and software
- vt_web hosting
- vt_hosted business applications
- vt_blogs and personal sites
- vt_specialized
- vt_blogs web communications
- vt_refinement
- vt_financial services
- vt_investing
- vt_south africa
- vt_software downloads
- vt_toolbars
- vt_project hosting
- vt_financial data and services
- vt_in
- vt_被劫域名
- vt_hacking
- vt_tree
- vt_technical analysis
- vt_brokerages trading
- vt_educational materials
- vt_not recommended site
- vt_sex
- vt_administration_and_school_management
- vt_porn
- vt_adult mature content
- vt_pornography
- vt_beer to peer
- vt_financial
- vt_known infection source
- vt_file sharing
- vt_education
- vt_real estate
- vt_rupee v directorios
- vt_Независимость
- vt_podcasts
- vt_media file download
- vt_educational institutions
- vt_ohio state university the
- vt_universiti kebajikan malaysia
- vt_iverson university
- vt_universidad politecnica de madrid
- vt_travel
- vt_personal network storage and backup
- vt_consolidators
- vt_personal network storage.ringtones mobilephone downloads software downloads
- vt_chains
- vt_hobbies
- vt_reference materials
- vt_cnk dl
- vt_reference
- vt_government
- vt_government legal
- vt_organizational email

VirusTotal

- fg_news and media
- fg_sports
- fg_entertainment
- fg_streaming media and download
- fg_proxy avoidance
- fg_search engines and portals
- fg_web analytics
- fg_personal vehicles
- fg_instant messaging
- fg_health and wellness
- fg_shopping
- fg_job search
- fg_freeware and software downloads
- fg_games
- fg_business
- fg_newsgroups and message boards
- fg_web-based email
- fg_web-based applications
- fg_information technology
- fg_pornography
- fg_finance and banking
- fg_brokerage and trading
- fg_personal websites and blogs
- fg_illegal or unethical
- fg_education
- fg_internet radio and tv
- fg_file sharing and storage
- fg_real estate
- fg_travel
- fg_reference
- fg_government and legal organizations

Fortiguard

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Methodology

**Empirical
validation**

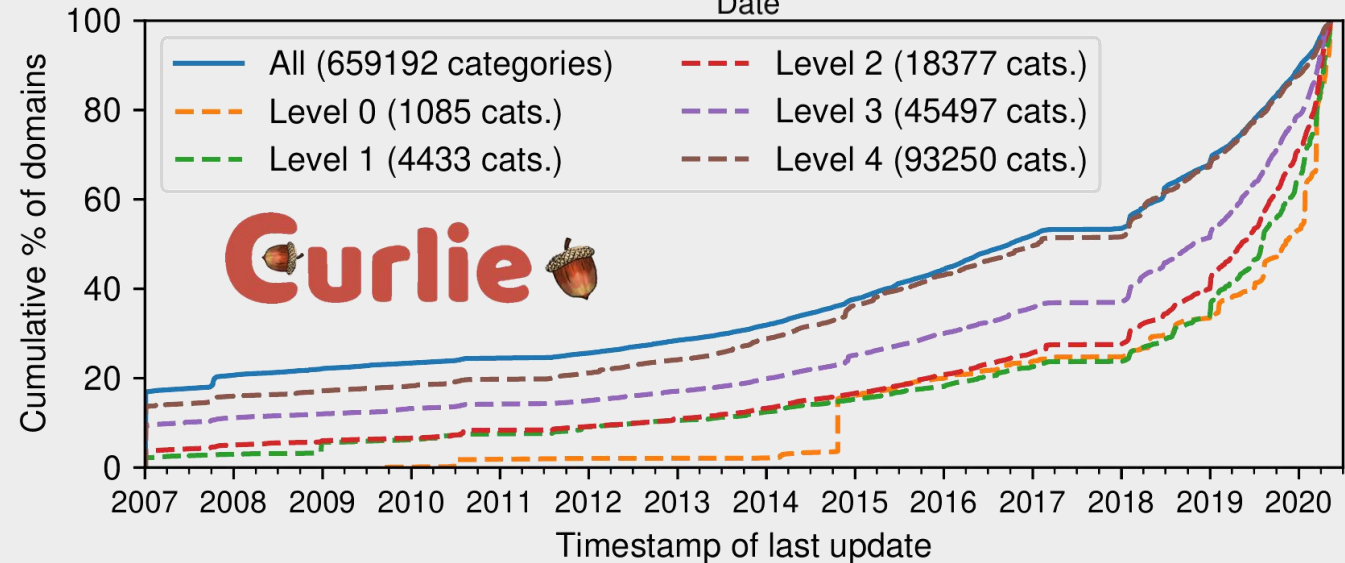
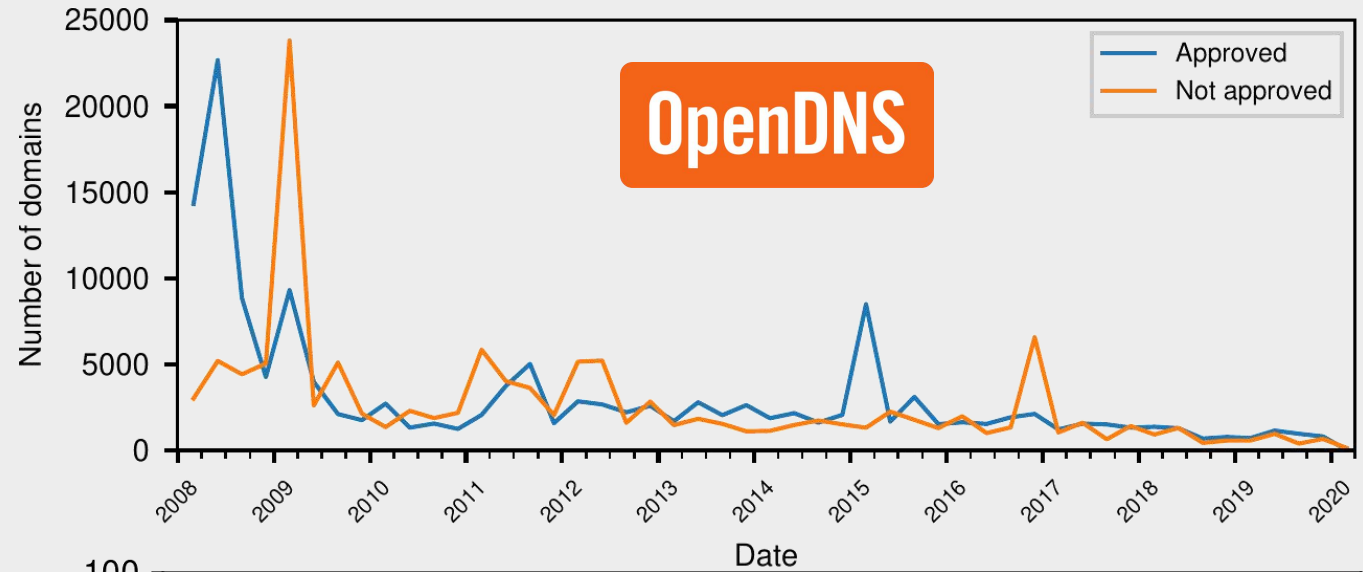
**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Dynamics of human labeling may trigger biases

Participation concentrated

- › at *beginning* of project
 - ›› **outdated** labels?
- › with *few* users
 - ›› **lack** of peer **review**?
- › on *unlabeled* domains
 - ›› **stale** labels?

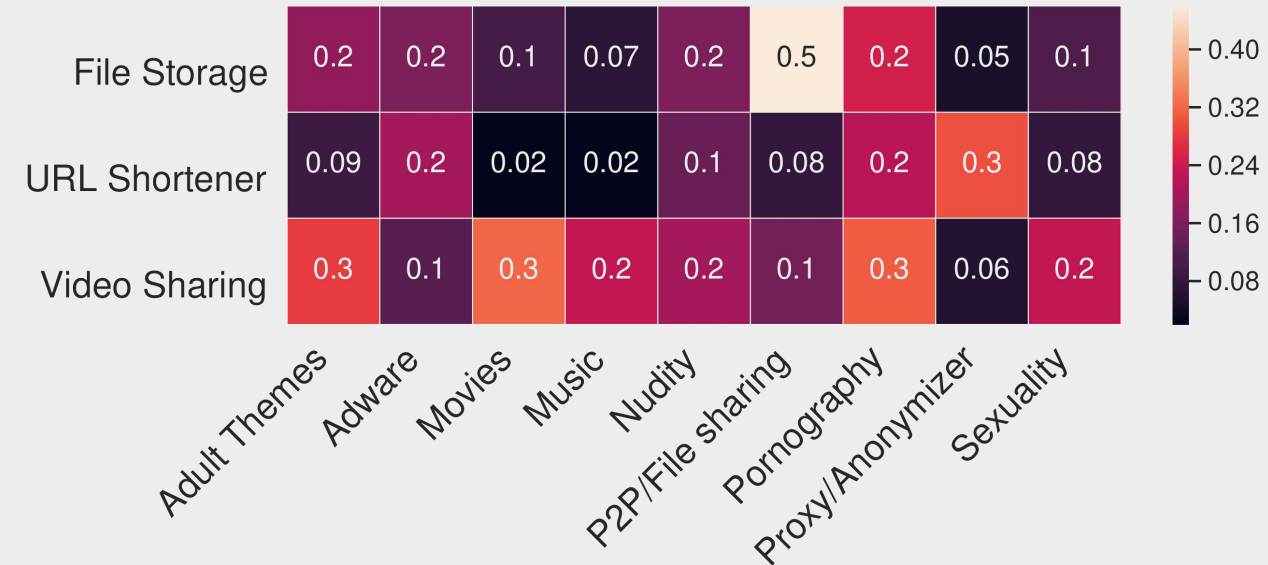
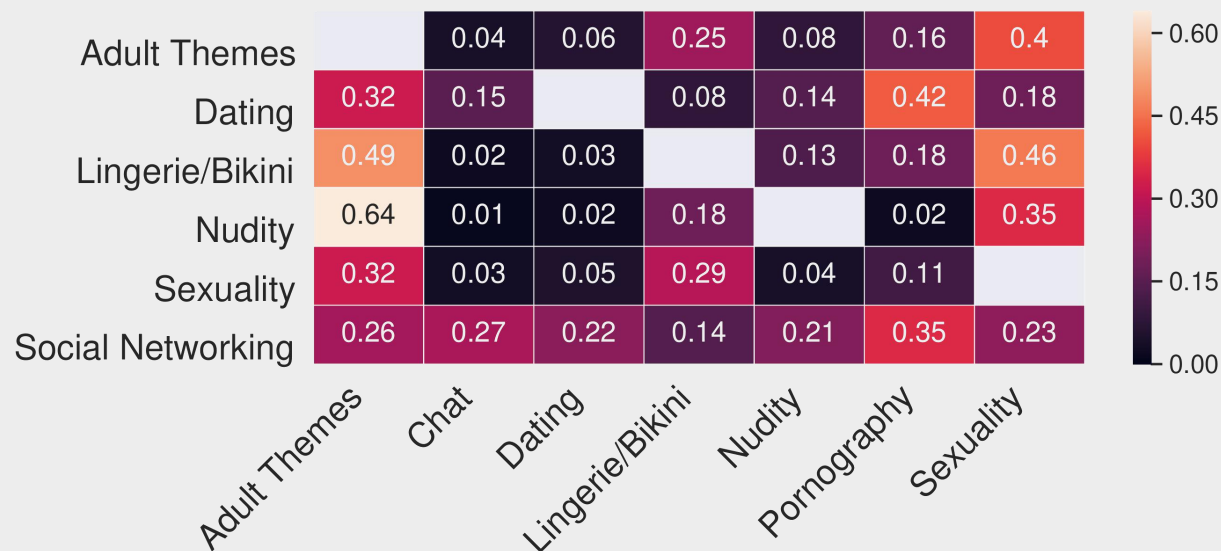


Disagreement in human labeling may trigger biases

- › Label assignment is **not completely objective**

Disagreement in human labeling may trigger biases

- › Label assignment is not completely objective
- › *Empirically*: Clusters of **correlated** labels



Disagreement in human labeling may trigger biases

- › Label assignment is not completely objective
- › *Empirically*: Clusters of correlated labels
- › *Experimentally*: 35.5% **disagreement** among authors,
71% matches community label

We analyze services on *specialized use cases*

- › Intended usage → requirements → data source selection
- › Service selection → characteristics → coverage/accuracy
- › Estimate **suitability** for three case studies
 - ›› Obtain a manually curated list as “*ground truth*”
 - ›› Analyze *coverage* across domains
 - ›› Analyze *appropriateness* of labels

Behavior differs widely for specialized use cases

› Advertising and tracking

›› *Curated list*: EasyList/EasyPrivacy

›› *Finding*: few services label the domains **at all**, let alone as tracker

› Adult content

›› *Curated list*: [Val19] and gambling regulators

›› *Finding*: 5 services label **correctly**, 3 others **hardly** label any

› CDNs/hosting providers

›› *Curated list*: signatures from WebPageTest

›› *Finding*: **confusion** between service *function* and *content*

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Recommendations

- › We avoid recommending a *specific* service
 - ›› “Best” service depends on *use case* and *requirements*
 - ›› We cannot measure *semantic agreement* nor *correctness*
- › Our recommendations address *best practices*

Recommendations

- › *Coverage and accuracy* may be **insufficient**
 - › Very *service-* and *use case-*dependent
 - › Consider **impact of errors**
- › *Purpose and updates* may introduce **biases**
 - › **Consult documentation** for *taxonomy* and *label sources*
 - › ... but **verify (and report)** manually, as **inconsistencies** exist
- › *Taxonomies* **differ** in size, scope and semantics
 - › Sound **aggregation** is **not obvious**

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Methodology

**Empirical
validation**

**Deep dive:
human labeling
& case studies**

**Discussion
&
Conclusion**

Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services

Pelayo Vallina, Victor Le Pochat, Álvaro Feal,
Marius Paraschiv, Julien Gamba, Tim Burke,
Oliver Hohlfeld, Juan Tapiador, Narseo Vallina-Rodriguez



References

- › [Yan04] Hsin-Chang Yang and Chung-Hong Lee. 2004. A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications* 27, 4 (2004), 645–663. <https://doi.org/10.1016/j.eswa.2004.06.009>
- › [Qi09] Xiaoguang Qi and Brian D. Davison. 2009. Web Page Classification: Features and Algorithms. *Comput. Surveys* 41, 2, Article 12 (Feb. 2009), 31 pages. <https://doi.org/10.1145/1459352.1459357>
- › [Bru20] Renato Bruni and Gianpiero Bianchi. 2020. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. *Expert Systems with Applications* 142, Article 113001 (2020), 14 pages. <https://doi.org/10.1016/j.eswa.2019.113001>
- › [Res04] Paul J. Resnick, Derek L. Hansen, and Caroline R. Richardson. 2004. Calculating Error Rates for Filtering Software. *Commun. ACM* 47, 9 (Sept. 2004), 67–71. <https://doi.org/10.1145/1015864.1015865>
- › [Ric02] Caroline R. Richardson, Paul J. Resnick, Derek L. Hansen, Holly A. Derry, and Victoria J. Rideout. 2002. Does Pornography-Blocking Software Block Access to Health Information on the Internet? *JAMA* 288, 22 (Dec. 2002), 2887–2894. <https://doi.org/10.1001/jama.288.22.2887>
- › [Sch18] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *IMC '18*. 478–493. <https://doi.org/10.1145/3278532.3278574>
- › [LeP19] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *NDSS '19*. 15. <https://doi.org/10.14722/ndss.2019.23386>
- › [Ahm20] Syed Suleman Ahmad, Muhammad Daniyal Dar, Muhammad Fareed Zaffar, Narseo Vallina-Rodriguez, and Rishab Nithyanand. 2020. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. In *WWW '20*. 271–280. <https://doi.org/10.1145/3366423.3380113>
- › [Zeb20] David Zeber, Sarah Bird, Camila Oliveira, Walter Rudametkin, Ilana Segall, Fredrik Wollmén, and Martin Lopatka. 2020. The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing. In *WWW '20*. 167–178. <https://doi.org/10.1145/3366423.3380104>
- › [Val19] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the porn: A comprehensive privacy analysis of the web porn ecosystem. In *IMC '19*.
- › [Seb16] Marcos Sebastián, Richard Rivera, Platon Kotzias, and Juan Caballero. 2016. AV-class: A Tool for Massive Malware Labeling. In *RAID '16*. 230–253. https://doi.org/10.1007/978-3-319-45719-2_11
- › [Lee13] Jung-Hyun Lee, Jongwoo Ha, Jin-Yong Jung, and Sangkeun Lee. 2013. Semantic Contextual Advertising Based on the Open Directory Project. *ACM Transactions on the Web* 7, 4, Article 24 (Nov. 2013), 22 pages. <https://doi.org/10.1145/2529995.2529997>
- › [Wei19] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *CCS '19*. 149–166. <https://doi.org/10.1145/3319535.3363200>