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

Pelayo Vallina
IMDEA Networks Institute /
Universidad Carlos III de Madrid

Victor Le Pochat
imec-DistriNet, KU Leuven

Álvaro Feal
IMDEA Networks Institute /
Universidad Carlos III de Madrid

Marius Paraschiv
IMDEA Networks Institute

Julien Gamba
IMDEA Networks Institute /
Universidad Carlos III de Madrid

Tim Burke
imec-DistriNet, KU Leuven

Oliver Hohlfeld
Brandenburg University of Technology

Juan Tapiador
Universidad Carlos III de Madrid

Narseo Vallina-Rodriguez
IMDEA Networks Institute/ICSI

ABSTRACT

Domain classification services have applications in multiple areas, including cybersecurity, content blocking, and targeted advertising. Yet, these services are often a black box in terms of their methodology to classifying domains, which makes it difficult to assess their strengths, aptness for specific applications, and limitations. In this work, we perform a large-scale analysis of 13 popular domain classification services on more than 4.4M hostnames. Our study empirically explores their methodologies, scalability limitations, label constellations, and their suitability to academic research as well as other practical applications such as content filtering. We find that the coverage varies enormously across providers, ranging from over 90% to below 1%. All services deviate from their documented taxonomy, hampering sound usage for research. Further, labels are highly inconsistent across providers, who show little agreement over domains, making it difficult to compare or combine these services. We also show how the dynamics of crowd-sourced efforts may be obstructed by scalability and coverage aspects as well as subjective disagreements among human labelers. Finally, through case studies, we showcase that most services are not fit for detecting specialized content for research or content-blocking purposes. We conclude with actionable recommendations on their usage based on our empirical insights and experience. Particularly, we focus on how users should handle the significant disparities observed across services both in technical solutions and in research.

CCS CONCEPTS

• **Networks → Network measurement**; • **Information systems → Clustering and classification**; *Web applications*; *Web searching and information discovery.*

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
IMC '20, October 27–29, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8138-3/20/10…$15.00
https://doi.org/10.1145/3419394.3423660

ACM Reference Format:
Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. 2020. Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3419394.3423660

1 INTRODUCTION

The need to classify websites became apparent in the early days of the Web. The first generation of domain classification services appeared in the late 1990s in the form of web directories. Notable examples from this period are Yahoo! Directory [1] and DMOZ[1] [3]. The main purpose of such services was to facilitate the discovery of web pages relevant to a certain topic of interest. To this end, human editors manually classified sites—often relying on suggested categories submitted by other users—into a purpose-specific taxonomy [4]. The quick expansion of the Internet soon put this approach to an end and led to the development of automated classification solutions [5–9].

As the Web grew in size, content, and applications, domain classification services became a valuable facilitator in multiple areas. One key application is traffic filtering, *i.e.,* networking solutions designed to block access to sites that are deemed dangerous (*e.g.,* phishing or malware [10, 11]) or inappropriate (*e.g.,* adult content). Cybersecurity firms such as McAfee [12] and OpenDNS [13] (Cisco) rapidly developed their own products. These technologies are nowadays embedded in multiple applications and setups such as parental control solutions and traffic filters in schools [14], libraries, and enterprise networks [15–17]. The online marketing industry also found domain classification extremely useful, in particular to improve targeted contextual advertising [18–20]. This led the Interactive Advertising Bureau (IAB) to develop an open standardized taxonomy for real-time bidding protocols [21]. Finally, networking, privacy, and security researchers also rely on website classification services to conduct category-dependent measurements [22–24] or to discover websites falling in a given category [25–27].

[1] DMOZ was closed in 2017 by its operator AOL. It has been continued by the Curlie project, which is still operating as of this writing [2].

To the best of our knowledge, no study so far has specifically analyzed the coverage, labels and applicability of domain classification services in different scenarios and research domains. Classifiers that were developed for different target applications or with different methodological approaches often exhibit disparate characteristics in terms of their coverage and taxonomies. This may have a substantial impact on how much the applications and studies that rely on them can be trusted. In fact, previous research studies reported the need for manual classification of websites due to the shortcomings of commercial services [9, 28, 29].

Unfortunately, the evaluation of these services is complicated by their opacity. While many services claim to apply machine learning algorithms, it is unclear how thoroughly they perform concrete analyses to validate their solutions, how comprehensive the underlying training data is, and, ultimately, how trustworthy and accurate the resulting classification is. Similarly, services such as DMOZ and OpenDNS that rely on human volunteers may be biased due to subjective opinions in the moderation process. Therefore, classification services may not succeed at adequately covering the large diversity of websites in both number and nature.

In this paper, we address the questions above by presenting a first analysis of domain classification services. Specifically, we make the following contributions:

- We analyze 13 popular services selected through purpose-specific web searches as well as through a survey of all the academic works published during 2019 (§ 2). We find that the results of 24 academic papers published in 9 relevant conferences (*e.g.,* IMC, WWW) depend on the outcome of the domain classification services that they use. Then, we present a qualitative analysis of the approach followed by these domain classification services according to their documentation. We find that key differences in their approaches might affect coverage and accuracy (§ 3).
- We evaluate the coverage of these services for both popular and unpopular domains, their labeling methodology, their taxonomies, and the (dis-)agreements across services when labeling the same domains. We crawl the labels assigned to 4.4M domains and find that most services lack coverage (only two services have a coverage above 55%), especially for non-popular domains. Furthermore, we show that their complex taxonomies (in particular for marketing-oriented classification services, with sometimes over 7.5k observed labels) hinder sound interpretation (§ 4).
- We study how introducing humans in the labeling process might impact the coverage and label consistency of those services (§ 5). We find that manual classification is affected by disagreements, ambiguities, and mismatches in the labeling process as well as biases in the distribution of users that submit votes and the workload of editors. This translates in some domains receiving as many as 58 rejected labels. To gain a better understanding of these challenges, we run a controlled experiment involving manual domain labeling and find disagreements in 35.5% of the cases.
- We explore the performance of domain classification services as tools to identify websites of interest. To do so, we run three case studies in the areas of detecting (and filtering) advertisement and tracking, adult content, and CDN or hosting infrastructure (§ 6). We find that the accuracy and coverage of the studied services is extremely low, and that the choice of one service or another significantly affects the outcome because of differences in coverage, which ranges from over 95% to below 1%.
- Finally, we discuss the implications of our findings for both the technical and academic applications of these services (§ 7). We also provide recommendations on how users should handle the significant disparities observed across services and identify a number of research questions for future work.

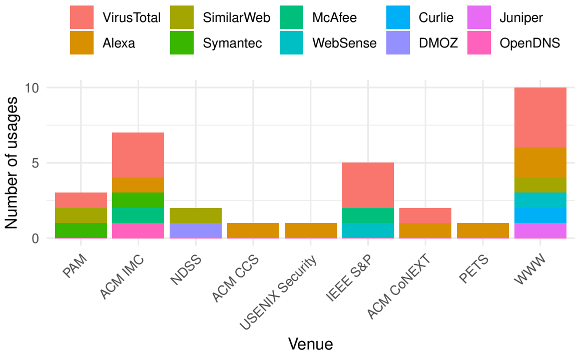


Figure 1: Usage of domain classification services in research during 2019. We have not observed the use of these services in TMA and ACM SIGCOMM papers in 2019.

2 USAGE IN ACADEMIC STUDIES

In this section we assess the relevance of domain classification services to academic studies. Given that the unknown properties of these services can impact research results, it is important to understand how widespread their usage is and what they are used for in the literature.

Survey approach. We survey all 1,014 papers published in 2019 at top venues in four areas: *i)* network measurements (IMC, PAM, TMA, CoNEXT, SIGCOMM); *ii)* security and privacy (CCS, NDSS, S&P, USENIX Security, PETS); and *iii)* Web (WWW). We first search for the names of domain classification services as well as keywords that indicate that such a service is used.[2] We then discard obvious false positives, such as the Amazon Alexa voice assistant instead of the Alexa domain classification service.

Usage. We manually analyze the remaining papers and find 26 papers that use at least one domain classification service (Figure 1). We find that for 24 (92%) of these, their results depend on the choice of service as they use it to gather their initial dataset or validate their results. Papers accepted at WWW and IMC are the ones that tend to rely the most on domain classification services. VirusTotal is the most popular service among academic studies (12 papers). Specifically, 3 papers [30–32] use the aggregate of VirusTotal's categories while 3 others [24, 26, 33] select one or more of the specific providers integrated in this popular threat-intelligence service. The remaining 6 papers [34–39] only rely on VirusTotal's detection of malicious domains or files. The second most popular service is Alexa, with 7 papers relying on it. All of these papers use

[2] The keywords used are "website classification", "website categorization", "domain categorization", "categorization service", "website category", "domain category", "category of the website", and "category of the domain", in singular and in plural, and also using British English spelling.

Alexa's lists of top sites per category to gather a corpus of websites (*e.g.,* governmental [40] or gambling and dating websites [26]). One paper [41] also uses the list of top sites per country. Our analysis reveals one paper using SurfControl [42], but as this service was acquired by Websense in 2007 [43], we do not consider it further. Table 6 in Appendix A lists all analyzed publications per venue.

Purpose. The 26 papers using domain classification services do so for a wide range of purposes. We find that 9 (35%) of them focus on security topics, including mobile sensors attacks [39] and certifications in the online payment industry [44]. We find 4 (15%) papers studying privacy in specific website categories—*e.g.,* tracking on pornographic websites [25]—or email tracking [31]. We identify 6 (23%) measurement papers, *e.g.,* on resource reloading by third-party websites [24] or web complaints [32]. Finally, 4 papers *question* the accuracy and applicability of existing domain classification services and either choose not to rely on them [28, 29] or manually validate the results [45, 46].

Takeaway: *We find that 26 papers published at top peer-reviewed conferences from 2019 use domain classification services. For 92% of these, their results depend on the choice of service, even though these services are sometimes questioned. As we will show later, in the absence of ground truth this dependence can introduce biases in the study results.*

3 METHODOLOGY OF DOMAIN CLASSIFICATION SERVICES

We perform an analysis of the 13 domain classification services listed in Table 1 using publicly available information. We select them based on their usage in recent academic works (§ 2), extending the set with services found through targeted online searches. Note that 2 of the domain classification services that we consider (FortiGuard and Webshrinker) were not used by any of the surveyed academic papers published in 2019. Our list does not cover all commercially available services, but those omitted pose a high barrier for data collection because of technical or monetary reasons.[3] Furthermore, VirusTotal is unique in that it does not provide its own classification, but instead aggregates category labels from third-party scanners. At the time of our data collection, these scanners were Alexa, Bitdefender, Dr.Web, Forcepoint, Trend Micro, and Websense.[4] However, since July 2020, these consist of (at least) Bitdefender, Comodo Valkyrie Verdict, Dr.Web, Forcepoint ThreatSeeker, Sophos, and Yandex Safebrowsing. We consider the former services (independently) in our evaluation. In § 4.2, we evaluate the consistency of services across multiple available sources.

Our evaluation focuses on features and methodological aspects that might affect how these services can be used in technical solutions and academic studies. Table 1 shows the features exhibited by the selected services according to their documentation and websites. A more detailed description of each service is provided in Appendix B. We also register our own domain and set up a live website hosting a WordPress blog, and then request its classification from each provider to investigate their approach to classifying new domains. We consider the following properties:

Table 1: Features of the analyzed classification services. For the services aggregated in VirusTotal, we list their properties as if they were accessed directly.

		Input		Output	Purpose				Updates			Access	
Service		Domain	Subdomain	Multiple categories	Content filtering	Threat assessment	Marketing	Discovery	Automated	Real-time	Reclassification	Free (sample)	Documentation
OpenDNS	[13]	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	[48]
McAfee	[12]	✓	✓	✓	✓	✓	✗	✗	✓✗	✗	✓	✓	[49]
FortiGuard	[50]	✓	✓	✗	✓	✗	✗	✗	✓✗	✗	✓	✓	[51]
VirusTotal	[52]	✓	✓	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
Alexa	[53]	✓✗	✓✗	✓	✗	✗	✓	✓	✗	✗	✓	✓	[54]
Bitdefender	[10]	–	–	–	✓	–	✗	✗	✓	–	✗	✗	[55]
Forcepoint	[56]	✓	✓	✗	✓	✓	✗	✗	✓✗	✓	✓	✓	[57]
Dr.Web	[58]	–	–	–	✓	✓	✗	✗	✓	–	✗	✗	✗
Trend Micro	[59]	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	[60]
Symantec	[11]	✓	✓	✓	✓	✗	✗	✗	✓✗	✓	✓	✓	[61]
Webshrinker	[19]	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	[62]
DMOZ	[3]	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓	[3]
Curlie	[2]	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓	[2]

Inputs. The granularity of input provided to the classifier affects the correctness of the classification: a subdomain may host a different kind of content than its base domain. For example, subdomains of the base domain (`yahoo.com`) may host a search engine (`search.yahoo.com`), a sports news site (`sports.yahoo.com`), or a webmail service (`mail.yahoo.com`). Depending on the origin of domains to be classified, *e.g.,* domain top lists often used by researchers that can include subdomains [63, 64], this can impact the accuracy and perception of the labels. All evaluated services may provide a separate classification for a subdomain. However, Alexa does not have a way to retrieve the classification given a (sub)domain. Instead, it requires searching through its listings of the top 500 domains in one of 279,716 categories.

Outputs. The outputs affect the utility of the data to a study's purpose. If a service yields multiple categories for a given site, this may improve the applicability and correctness of the classification as it can be more nuanced, *e.g.,* tagging a sports news website as both *sports* and *news.* However, this could also lead to an incoherent interpretation, *e.g.,* double-counting when aggregating domains by category. All services except FortiGuard and Forcepoint can assign multiple categories to domains.

Purpose. In many cases, the provider's intended purpose for a service (*e.g.,* content filtering, threat protection, marketing or discovery of relevant content) influences the used taxonomy. For example, a content-filtering service may prefer to label `youtube.com` purely as a bandwidth-consuming site, but a marketing-oriented service may label it as a video sharing or advertising platform. Most of the classification services analyzed are intended for content filtering, usually being integrated into their consumer or business web security software. One exception is VirusTotal, which provides only a threat assessment. Further exceptions are Alexa, DMOZ, and Curlie, which are designed for discovering sites within categories of interest. Moreover, certain services also have other applications. For instance, Webshrinker can categorize domains according to

[3] *e.g.,* Zvelo and Cyren require completing a reCAPTCHA for every request.

[4] Websense renamed itself to Forcepoint [47] after the acquisition of Stonesoft, yet both are listed separately in VirusTotal.

the marketing-oriented taxonomy of the Interactive Advertising Bureau (IAB) [21].

Updates. The ability to update classification results affects both coverage and accuracy. Real-time classification, often enabled by a fully automated analysis, may improve coverage and maintain data relevance. In other words, new sites can be immediately assigned to a category, and the classification will reflect the most recent content. For example, a change in website ownership would not result in outdated labels. Automated approaches may also increase the scale at which domains can be classified, in particular when additional data is used to label uncrawlable domains (*e.g.,* malware domains). The ability to request reclassification of a site may allow to correct errors, but it may also be leveraged to undeservedly receive a less "harmful" classification if requests are not adequately reviewed. For example, an adult website may attempt to get reclassified as a (non-adult) video streaming site in order to evade filtering.

Only Forcepoint, Symantec and Webshrinker provide real-time results: we confirm through web server logs that upon request, they immediately visit and categorize a domain that we newly registered. Webshrinker even proactively visits the domain (likely due to its entry in the zone file), and is the only one to deploy a real browser. This behavior can be traced back to the methods that services claim to use, mostly consisting of automated classification through machine learning algorithms. McAfee [49], FortiGuard [65], Bitdefender [66], Forcepoint [57], Symantec [67, 68], and Webshrinker [69] state in their documentation that they complement their crawler-based ML solution with domain metadata, security honeypots and scanners, and third-party feeds and logs, as well as human reviewers who inspect and amend automatically determined categories. OpenDNS, DMOZ, and Curlie rely on human volunteers to propose and confirm categories; Alexa uses a truncated version of DMOZ's data and taxonomy [54]. All services except VirusTotal, Bitdefender and Dr.Web provide a way to request domain reclassification: for our newly registered domain, the delay of several days before any change suggests that this process requires human intervention.

Access. Easy access to data and documentation improves usability for end users and researchers. For instance, clear descriptions and examples of sites that are considered part of a category aid in selecting the appropriate categories for other websites. Bitdefender and Dr.Web do not provide direct free access to their data, but they are available through VirusTotal. Dr.Web is the only service that does not document its taxonomy. VirusTotal does not document where and how it sources its data. In § 4.3, we compare the documented categories with those that we observe empirically.

Takeaway: *The substantial differences in domain classification services' characteristics affect their applicability: label interpretation depends on a service's supported inputs and outputs as well as taxonomy differences due to their purpose, while coverage and accuracy benefit from easy access to up-to-date labels. These properties should therefore be well understood to ensure correct application. We assess the veracity of services' claims through our own empirical observations in § 4, to determine their effective suitability to different scenarios.*

4 DOMAIN LABELING QUALITY

In this section we analyze domain classification services on their labeling coverage (§ 4.2), their individual taxonomies (§ 4.3), and the labeling consistency and relationships across providers (§ 4.4). In this analysis, we omit DMOZ and Curlie as they aspire to achieve a different goal, *i.e.,* supporting content discovery instead of concisely classifying all domains. This affects their data retrieval strategy and interpretation, and we would need to reverse their mapping of deeply nested categories to relevant domains.

4.1 Data collection

Our data collection process consists of two stages:

(1) **Compiling target domains.** We compile a large list of domains starting from the union of all daily Alexa top sites rankings between September 1 and 30, 2019. To reduce possible biases caused by the instability of the Alexa ranking [22, 63, 64], we aggregate these rankings using the default method of the Tranco top list [64], which sums domain scores from individual lists following a Zipf-like distribution. We retain a ranked list of 4,424,142 domains that we could successfully collect from all non-rate-limited services. While these 4.4M domains represent a small fraction of all registered domains [72], they are considered to be popular by the Alexa traffic ranking service. Their popularity is further reflected by the fact that 47% of the 4.4M domains are indexed in the Chrome User Experience Report [70] and 0.5% by Common Crawl [71], both generated between August and October 2019. We therefore believe that our set is representative of domains regularly visited by end users and therefore also of interest to researchers.

(2) **Crawling domain classification services.** We retrieve the category labels for the 11 selected domain classification services. As each service differs in how its online portals retrieve data, we develop the most scalable and least resource-intensive method possible for each provider.

- For FortiGuard, McAfee, and OpenDNS, we retrieve labels through their publicly available portals. While these services are not rate-limited and their data is public, we perform our data collection at a non-intensive average rate of 40 requests per minute. We retrieve McAfee's labels for its "Real-Time Database" product. For VirusTotal, we retrieve labels through its API, which aggregates six services: Alexa, Bitdefender, Dr.Web, Forcepoint, Trend Micro, and Websense. We received access to VirusTotal's academic API, with a request limit of 20k queries per day and account.
- For Symantec, Trend Micro and Webshrinker, our data collection is subject to rate limiting. Therefore, we retrieve labels on these three services for the top-10k domains in our ranked list. We retrieve Webshrinker's labels from its default marketing-oriented IAB taxonomy.

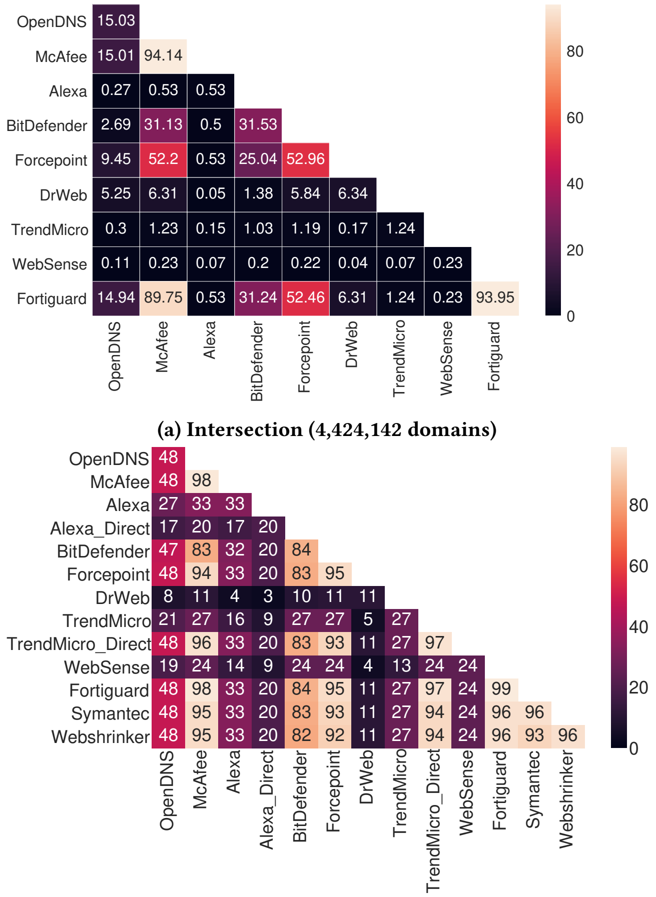
4.2 Coverage

One critical aspect to consider when using domain classification services is their coverage, defined as the number of websites for which they provide a meaningful label. This metric affects how comprehensively a service can both execute its original task and be deployed for large-scale applications and studies. As discussed in § 3, some domain classification services involve humans in the loop, while others try to achieve a larger scale or real-time classification using machine learning methods. As a result, not all services have the same ability to scale their labeling process. When measuring coverage, we apply a sanitization process to address the fact that five services (FortiGuard, OpenDNS, Websense, Forcepoint and Trend Micro) provide explicit labels for unclassified domains. We consider a domain "unlabeled" if we obtain an empty result, or a label explicitly stating that the service has not (yet) labeled the domain (e.g. *Uncategorized* for Forcepoint).

Figure 2a shows for which percentage of our full set of 4.4M domains we obtain a valid label. The diagonal reveals that the coverage varies greatly between *individual services*. The off-diagonal values report the 'intersection coverage' defined as the number of domains that both services label simultaneously, regardless of the label provided. FortiGuard and McAfee excel by labeling around 94% of domains, likely due to their deployment of machine learning techniques for automated classification. Contrarily, OpenDNS only achieves 15% coverage, with its manual submission and voting processes (§ 5) likely becoming a bottleneck when dealing with the millions of monthly domain registrations [72]. Alexa's coverage is even lower at 0.53%, possibly due to its data source DMOZ [54] containing human-volunteered labels in often highly specialized (and therefore less popular) categories designed for content discovery, as well as its limit of 500 websites per category. Services retrieved through VirusTotal also have much lower coverage; we will show later on that this may in part reflect a service integration issue at VirusTotal, as services do yield a label when directly queried.

For completeness, we also compute the "union coverage" between pairs of providers. We define it as the percentage of websites for which at least one service provides a valid label (Appendix C). This analysis suggests that considering the union of two services does not necessarily increase the global coverage when their intersection is already high. For example, the union coverage for FortiGuard and McAfee increases slightly to over 98%. However, as we will discuss in § 7, the combination of labels from multiple services is non-trivial due to largely disjoint taxonomies. As a result, unless the objective of unifying providers is offering complementary perspectives, it might not necessarily benefit coverage.

The importance of being popular. Table 2 shows that service coverage differs depending on domain popularity. We expect automated services to achieve a higher coverage even for less popular domains, but we observe that while McAfee and FortiGuard maintain a consistent coverage of at least 93% throughout, Bitdefender and Forcepoint drop from 93% and 98% to 27% and 48%, respectively, when labeling domains from either the top-1k or unpopular domains found in the long tail over 1M. We observe a similar behavior for Dr.Web, Websense, Trend Micro, and Alexa, who have relatively low coverage overall but perform worse for non-popular websites. The human labeling efforts of OpenDNS appear to prioritize popular domains (an expected feature). Nevertheless, OpenDNS coverage across domains ranked over the top-1M may be inflated by the 15% subdomains within that interval. As we will discuss next, in OpenDNS, subdomains typically inherit the label of the base domain. Finally, Trend Micro (directly sourced), Symantec and Webshrinker achieve a very high coverage of over 96% for the top-10k, but their rate limits make large-scale data collection unfeasible. In summary, only two services are able to categorize both popular and non-popular domains. Given the ever-increasing number of websites as well as the trend to conduct large-scale measurements, the choice of service impacts the capacity to classify potentially millions of visited or targeted domains, including undesired ones.



(a) Intersection (4,424,142 domains)

(b) Intersection (Top-10k domains)

Figure 2: Coverage per service (diagonal) and intersection of the coverage between pairs of services for our two domain sets (§ 4.1).

Base domain vs. Subdomains. We identify 582,230 (13%) subdomains among our 4.4M domains. Three services—OpenDNS, McAfee, and FortiGuard—provide labels for more than 99% of them. Yet, as we will see in § 4.3, there is no difference between base and subdomain labels in the majority of cases. In the case of OpenDNS, the improvement compared to its overall coverage (15%) stems from its approach to labeling subdomains. When humans do not offer a category for a subdomain, OpenDNS classifies it by default with the label of the base domain (if labeled). However, this coverage is skewed towards the 77% subdomains related to three base domains: `blogspot.com`, `wordpress.com`, and `tumblr.com`. For Alexa, Websense, and Trend Micro, subdomain coverage is below 1%. Depending on the source and selection of domains, overall coverage may therefore become worse.

Direct Source vs. VirusTotal. We verify labels collected through VirusTotal (which aggregates 6 existing services) by directly collecting labels for the top-10k domains at two services, Trend Micro

Table 2: Coverage for different domain popularity intervals. For each interval, we list the number of domains for which we could successfully collect labels.

Rank	(0-1k]	(1k-10k]	(10k-100k]	(100k-1M]	+1M	Overall
# domains	1,000	8,945	89,276	678,246	3,646,675	4,424,142
OpenDNS	78%	46%	19%	9%	16%	15%
McAfee	100%	99%	98%	97%	94%	94%
FortiGuard	100%	100%	99%	97%	93%	94%
Alexa						
through VT*	48%	32%	13%	1.02%	0.05%	0.5%
direct source	31%	20%	-	-	-	-
Bitdefender*	93%	83%	73%	48%	27%	32%
Forcepoint*	98%	95%	90%	73%	48%	53%
Dr.Web*	16%	11%	6%	4.2%	7%	6%
Trend Micro						
through VT*	55%	25%	9%	2.7%	0.7%	1.2%
direct source	98%	97%	-	-	-	-
Websense*	52%	22%	3.9%	0.36%	0.04%	0.2%
Symantec	99%	96%	-	-	-	-
Webshrinker	98%	97%	-	-	-	-

*Retrieved through VirusTotal.

Table 3: Differences between the results obtained through direct sources vs. VirusTotal. For this comparison, we use the top 10,000 domains in our ranked list.

	Direct Source		VirusTotal		Intersection	
	Coverage	# Labels	Coverage	# Labels	# Labels	Consistency
Alexa	21%	1,843	33%	1,719	35	2.6%
Trend Micro	98%	75	27%	63	817	27%

and Alexa. As shown in Table 3, Trend Micro's coverage is much higher (98%) when directly queried than when using VirusTotal (28%). Moreover, only 27% of the domains are classified with the same label and only half of the distinct labels appear at both sources. As we will expand on in § 4.3, we suspect VirusTotal may be using a different or an older Trend Micro product, with a potentially lower coverage and different set of labels. However, for Alexa we observe the opposite behavior: we obtain 12% more coverage through VirusTotal. Again, this may point to VirusTotal obtaining Alexa's data from an unknown source, different to our (one-time) search within the top 500 sites of Alexa's 279,716 categories. The inconsistencies between VirusTotal and a direct source indicate that the former might not be a fully reliable source. This is particularly worrisome given VirusTotal's popularity in recent academic work (§ 2).

4.3 Labels within services

In this section, we report on the distinct labels that we observe in each service, and the properties that affect their correct and tractable interpretation: their diversity, deviations from documentation, and uniqueness. We normalize all labels to lowercase, and we break down multi-labeled classifications into their individual units to reduce possible inconsistencies in the comparison.

Label diversity. Table 4 shows that the number of observed labels per service varies significantly across services, but conforms to their intended purpose. Security and content filtering services have fewer labels (12 observed in Dr.Web to 125 observed/139 documented in Forcepoint), which may simplify the setup of security policies. Conversely, the larger diversity in marketing-oriented services (300 observed/401 documented in Webshrinker, and more than 7,500 observed in Alexa) may enable more fine-grained targeting. We also see that all services except Websense use at least one label that is unique to them, showing that their taxonomies are diverse and not trivial to merge. While some services offer hierarchical taxonomies that can reduce the diversity by replacing a label with that of an ancestor, this compromises precision and forces users to decide where to prune the tree. This complexity is best exemplified by labels for Alexa queried through VirusTotal, which will only yield the label of the leaf. This is often a non-English label, derived from that website's classification into the multilingual *World* tree. For example, a given domain may be labeled as *Arts* (English), *Artes* (Spanish), or *Kultur* (German). In short, it is hard to reduce the large set of labels, without affecting their usability and interpretability.

Table 4: Comparison of documented (*Doc.*) and observed (*Obs.*) labels, including labels unique to a particular service, across 4.4M analyzed domains unless otherwise stated.

Service	# Obs.	# Unique obs.	# Doc.	# Obs. not doc.	# Doc. not obs.
OpenDNS	64	26	58	5	0
McAfee	108	71	102	6	1
FortiGuard	87	42	86	1	0
Alexa*	7,557	7,417	–	–	–
Bitdefender*	60	34	43	25	9
Forcepoint*	125	18	139	3	21
Dr.Web*	12	6	–	–	–
Trend Micro					
through VT*					
2019 taxonomy	84	37	86	15	17
2011 taxonomy	84	37	84	7	7
direct source**					
2019 taxonomy	77	31	86	2	11
2011 taxonomy	77	31	84	9	16
Websense*	99	0	139	2	45
Symantec**	79	42	90	0	11
Webshrinker**	299	212	401	1	103

*Retrieved through VirusTotal.
**Across the top-10k domains in our ranked list. These counts are therefore lower/upper bounds of those across all 4.4M domains.

Documented vs. Observed labels. In order to further understand how well these services document their taxonomy, we compare the documented categories with those that we observe in our dataset. As shown in Table 4, we observe at least one undocumented category for every service except Symantec; while Alexa doesn't explicitly document its categories, we observe only 7,557 labels for Alexa through VirusTotal, far fewer than in our own search (279,716 categories). Certain differences are due to minor syntactical variations (*e.g.,* the documented *Non-traditional religions* versus the observed *Non-traditional religion_* in Forcepoint), yet they might affect researchers who search for a particular documented category and are unable to find sites within it. Other differences are due to potentially incomplete or outdated documentation. For McAfee, we still observe six categories that have been deprecated since 2010 according to their own documentation [49]. For OpenDNS, five security-related categories are unavailable for user submission or voting, as they either are restricted to trusted sources (*e.g., malware*), or appear to be legacy categories (*e.g., adware* [73]). For the Trend Micro data sourced from VirusTotal, there is a higher correspondence with its 2011 taxonomy [74] than with its 2019 one [60], suggesting that VirusTotal sources classifications from an older Trend Micro product. Finally, certain sensitive categories appear to be omitted from the documentation, *e.g., homosexuality* in FortiGuard. In summary, service documentation cannot be trusted to fully reflect the taxonomy observed in the wild, countering correct configuration and sound research usage.

Multilabeling. Six services (OpenDNS, McAfee, Dr.Web, Forcepoint, Trend Micro, and Websense) use multiple labels to categorize a single domain. This is uncommon behavior for most services, except in Dr.Web, where 67% of the domains have multiple categories, while the presence of multi-label domains is anecdotal in Forcepoint and Websense, at less than 1% of the labeled domains. Nevertheless, the number of labels that a domain can have varies for every service: in Trend Micro, 7% of multi-label domains have three or more labels, while there is only one such domain for Forcepoint. While for other services we observe at most 6 labels for one domain, in OpenDNS, we observe `4chan.org` reaching a maximum of 17 labels. Multiple labels may add nuance, but also complexity to their interpretation.

Next, we measure which pairs of labels frequently appear together. We observe 2,536, 1,006, 526 and 356 distinct pairs in McAfee, OpenDNS, Trend Micro and Forcepoint respectively. However, in Dr.Web and Websense, this number drops to 44 and 40; for the former, this is due to the low number of labels observed (Table 4). The label pairs are often unevenly distributed, *e.g.,* in Trend Micro, 2% of the labeled domains have the most popular pair *disease vector-spam*, while the next most popular pair *financial services-business economy* appears only on 0.2% of the domains. In McAfee and OpenDNS, the most popular pairs, *personal pages-internet services* and *blogs-content delivery networks*, appear on 1% and 39% of labeled domains respectively. Common pairs are also not always intuitively linked. For example, in Dr.Web, the most popular pair is *adult content-social network*, appearing in 65% of all domains labeled by Dr.Web, where 60% of them are subdomains of `blogspot.com`. When using aggregated labels from VirusTotal without taking into account individual services, a non-adult blog could, therefore, be inadvertently labeled as an adult site, impacting applications targeting adult content.

Base domain vs. Subdomains. We saw in the previous section that coverage on subdomains is better compared to the general coverage, in the case of OpenDNS with an improvement of 70%. We now analyze how meaningful these labels are. We see that for OpenDNS, McAfee, and FortiGuard, 99%, 98%, and 97% of subdomains, respectively, have at least the label of the base domain. However, since domains at McAfee and OpenDNS can be multi-labeled, we observe that the percentage of the subdomains that have the same labels as the base domain drops to 46% in OpenDNS, while in McAfee, below 1% of the subdomains have different labels. This drop in OpenDNS is because 90% of `blogspot.com` subdomains, which represent 51% of the total subdomains observed, have the original label of the base domain (*Blogs*) plus an extra label, typically *Content Delivery Networks* (90% of cases). We conclude that subdomains inherit the label of the base domain, without taking into account the actual content of the subdomain.

Labeling update. As discussed in § 3, the frequency of label updates affects the timeliness and, therefore, accuracy of labels. We analyze how common such updates are for the 9 services that do not rate limit (see § 3). We select 2,000 domains per service: half of them were previously labeled by (at least) that service, while the rest were unlabeled for the particular service. We select domains that have been crawled at the beginning of our data collection, to increase the time that these services had to (re-)label the domains.

We find that in our second round, only OpenDNS, FortiGuard, and McAfee categorize domains that had not been previously labeled. However, the number of updates varies: while McAfee and FortiGuard now label 88 and 53 out of 1,000 previously unlabeled domains, OpenDNS only does so for 2 domains. Similarly, for domains that had been previously labeled, McAfee and FortiGuard relabel 15 and 10 domains, respectively. The majority of these changes concern the maliciousness of domains, with some of them gaining a related label (*e.g., malicious sites*) while others lose such a label. Finally, for OpenDNS, three domains gain a label, although two of those receive the label *Content Delivery Networks* outside of the regular voting process (§ 5). In summary, some services update labels over time, making it more likely that their classification better reflects the current state of a website.

4.4 Labels across services

The differences in both label number and coverage (see Table 4) call for a better understanding of the relationships between services. This analysis is however hindered by inconsistencies in label syntax (*e.g., News* vs. *News and Media*), language (*e.g., Arts* vs. *Artes*), semantics (*e.g., File sharing* vs. *File storage*), and aggregation (*e.g., sports* vs. *entertainment/sports*). Furthermore, one provider may give multiple labels to a particular domain, requiring a comparison of sets of labels with different dimensions.

Mutual information. In this section, we take a statistical approach to perform a label-agnostic analysis. A suitable metric is the *mutual information*, which describes the amount of information gained about a random variable upon observing another random variable [75]. Mutual information can be thought of as the reduction in one variable's entropy (level of uncertainty) if the output of another variable is observed. In our case, we treat each provider as a random variable whose distribution of values (*i.e.,* labels) we estimate empirically. We can then interpret the mutual information as how similarly the labels are distributed between two services. Its normalized value will be 1 if one service assigns a common label to all domains (and none other) that are given a common label by the other, regardless of the exact label syntax. Conversely, it will be 0 if the services are completely independent, *i.e.,* there is no information to be gained about the first when observing the labels of the second.

We select McAfee, OpenDNS, Bitdefender, Forcepoint, VirusTotal and FortiGuard for this analysis as they are the services with the largest coverage (see Table 2). VirusTotal is a special case: while it meets the coverage criterion, its labels are aggregated from other providers, including Bitdefender and Forcepoint. The normalized mutual information matrix is shown in Figure 3. Overall values are

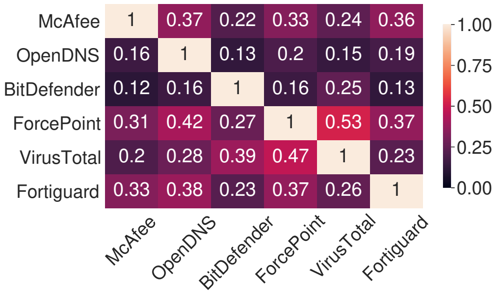


Figure 3: Normalized mutual information of domains with the highest degree of overlap.

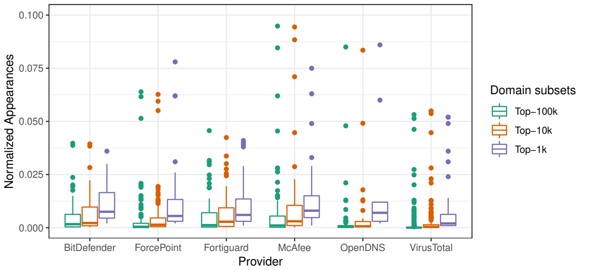


Figure 4: Normalized label occurrence frequencies. The statistics are computed over the number of times a label repeats itself for a given range of domains.

low, indicating disagreement between providers, which is due to several reasons. First, services such as OpenDNS and Bitdefender differ in specialization, providing either a content- or a security-oriented label, *e.g., Online Service* vs. *Spam*. Next, human-sourced services such as OpenDNS may suffer more from subjective labeling (§ 5) and therefore disagree more with automated services such as McAfee. Differences in the size and granularity of taxonomies (*e.g.,* between VirusTotal and FortiGuard) can introduce further disagreement. Finally, shared sources of labels or taxonomies may inflate agreement: we see the highest mutual information between VirusTotal and two of its aggregated providers, due to their partially shared data source. We observe consistent results when repeating our analysis using the conditional entropy.

Label frequency. Next, we compare the distribution of labels over domains, in order to understand the label coverage as well as service specialization. Figure 4 presents the normalized label frequencies for the top-1k, 10k and 100k domains in our ranked list. In all three subsets but in particular for the top-100k, there is a significant number of outlier labels that appear with a much higher frequency, indicating that labels are distributed unevenly. With the exception of VirusTotal, the median frequency for labels across domains is relatively consistent. On the top-1k domains, OpenDNS shows the smallest granularity in terms of coverage, while VirusTotal shows the highest. The trend is partially maintained when considering larger domain sets, where Bitdefender, FortiGuard and McAfee span the considered domains with the smallest number of labels.

Label distribution. Finally, we observe two trends in the concrete distributions of labels between providers. First, we see that, especially when considering more than two providers, one fixed set of domains corresponds to largely varying sets of labels that cannot trivially be combined into one category: *e.g., Nudity, Society and Lifestyle*, and *Adult Content* are overlapping but not equivalent categories. We provide a visual example of these inter-service label relationships in Appendix D. Secondly, we find that labels are distributed unevenly across pairs of providers: *e.g.,* for McAfee, the lower granularity of its taxonomy means that few labels cover the set of domains generated by a large number of labels from other services while for VirusTotal far more labels are needed. This distribution of labels across services is further explored in Appendix E. In summary, differences in service purpose, taxonomy size and label distribution cause large disagreements between services, making it difficult to compare and combine their classifications.

Takeaway: *We find that commonly used domain classification services exhibit traits that affect their suitability, both for technical solutions as well as for research. Only a few services attain a level of coverage that is sufficient to cover non-popular or non-base domains. Services may return multiple or undocumented labels, requiring careful data processing and even manual validation. Breaking down multi-labeled classification may ease the label comparison between services as well as improve the interpretation of the results. However, it may also bias the results, overestimating the presence of labels that do not provide information about the real purpose of the service. The large diversity in labels, both within and across services, may harm their accurate and tractable interpretation. Efforts to combine labels from multiple services to achieve a higher agreement on label accuracy might be thwarted by labeling inconsistencies. The labeling updates may also have an impact on accuracy and timeliness. Researchers should be aware of these phenomena and renew their dataset to reduce possible misclassifications, especially in treating malicious services. In summary, sound deployment and usage of domain classification services requires a thorough understanding of the (desired) characteristics and resulting biases to select the most appropriate sources.*

5 HUMAN PERCEPTIONS

As described in § 3, OpenDNS, DMOZ and Curlie leverage a network of human volunteers to label domains. In OpenDNS, moderators approve or reject labels voted on by users, while in DMOZ and Curlie, editors add suggested sites to their managed categories. In this section, we harvest historical data from OpenDNS' voting process to further measure the effect that human decisions have on (1) OpenDNS' labeling process—in terms of user and editor temporal dynamics—and, (2) on the resulting classifications. For comparison and completeness, we also study Curlie's labeling dynamics by crawling and analyzing their publicly available data.

5.1 Labeling dynamics

OpenDNS. OpenDNS relies on a voting process that allows users to submit labels ('tags') for domains, which then receive positive and negative votes from other users. After sufficient votes, a trusted moderator approves or rejects these submitted labels [76]. OpenDNS publicly releases historical data from this voting process,

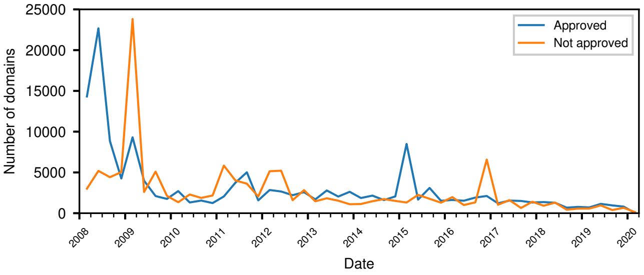


Figure 5: Domains labeled in OpenDNS by quarter.

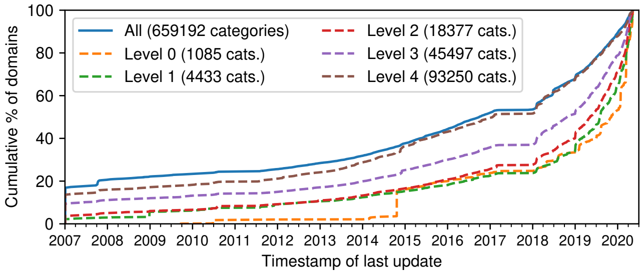


Figure 6: Cumulative distribution of update timestamps for categories in Curlie.

including the labels proposed for every domain, the user who proposed them, whether they are accepted or not, and the moderator who took the final decision. All items are timestamped, which allows us to analyze the evolution of submitted labels over time. This data allows us to inspect the OpenDNS voting process for 794.8k domains, as well as the behavior, agreements and disagreements between 19k users and 292 moderators from February 2008 until January 2020.

First, we analyze who is submitting labels for observed domains. The first observation that stands out is that most users are "casual," as 95% of users only submit a label for 10 domains or fewer. Nevertheless, there is a group of 160 highly engaged users who submitted labels for more than 100 domains. As for moderation, the workload distribution is more even: around 40% of moderators have approved 10 labels or fewer. Nevertheless, there are 292 moderators (0.03% of all moderators) which are very prolific, being responsible for the approval of over 10k labels.

Figure 5 shows the number of approved and not approved labels submitted quarterly. We can observe that the majority of labels were submitted during 2008 and 2009. Interestingly, at the beginning, the majority of labels were accepted. However, starting in 2009 there is a large decay on the number of accepted labels and an increment of those that are not accepted. Our intuition is that because at the beginning of the project all major sites lacked a label, the probability of people correctly labeling those is higher. As time passes, only a long tail of unpopular domains remain unlabeled, so users are more likely to submit an incorrect label or no label at all.

Curlie. As in the case of DMOZ, Curlie has no open voting process. Instead, trusted editors fully manage categories and decide which user suggestions they include. Review may come from other editors for the same category and its parent categories, or those with the right to edit all categories [77, 78]. Because of its content discovery purpose, Curlie has a large and deep hierarchical taxonomy, consisting of 671,715 observed categories. By analyzing the assignments of categories to editors, we examine whether these editing and reviewing processes can be effective considering this deep taxonomy.

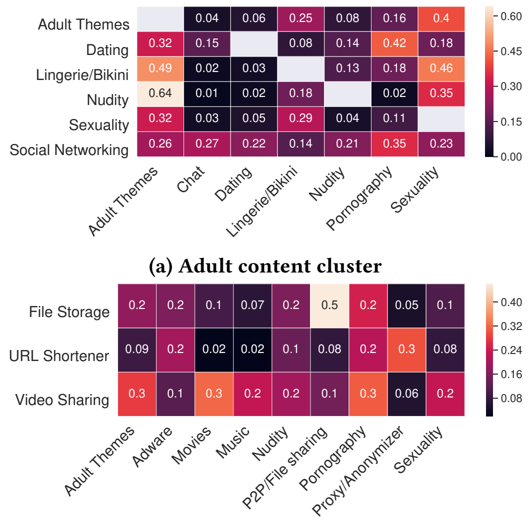
Only 985 (0.1%) are explicitly managed by at least one out of 294 active editors. When we account for the editing rights to subcategories, 515,791 (76.8%) categories have *at least* one "implicit" editor. However, 565,812 categories have *at most* one implicit editor, which means that 84.2% of categories can only be peer reviewed by the editors with rights to all categories. The opportunity for peer review may be further affected by the breadth of certain editors' scope, with the top "implicit" editor managing over 300k categories. In summary, the large number of categories managed by only a few editors may prevent these editors from conducting a regular review for accuracy and recency.

Figure 6 shows that around half of all categories have been updated since the evolution of DMOZ into Curlie in 2017. Moreover, it shows more recent activity higher in the tree: lower levels may either inherently require fewer updates, or may be less actively maintained by their editors. While there is steady ongoing activity on Curlie, many categories have not been updated for years, potentially leading to their entries being outdated or inaccurate.

5.2 Labeling (dis-)agreements

One key issue with human-in-the-loop labeling is that the task of classifying domains is not completely objective, and thus different users might suggest different labels for the same website. Therefore, we measure how often this happens in the labeling process of OpenDNS. While the median number of accepted and rejected labels in OpenDNS is one, we have shown in § 4.3 that some domains have as many as 17 accepted labels. In the case of labels that do not get approved, we can find domains with a high level of disagreement among voters with as many as 58 not accepted labels.

We further investigate the type of labels that create most agreement and disagreement in OpenDNS. To do so, for all domains with a given approved label, we measure how often other proposed labels are approved and rejected for the same domain. Selected clusters of labels where the disagreement is high are shown in Figure 7. Some of the labels that often appear together seem to be a product of honest mistakes by the users, as they are closely related (such as *Adult themes* and *Sexuality*, or *Travel* and *Business Services*). An interesting case is the label *Pornography*, which often appears proposed (and rejected) in addition to other labels. While this might make sense for some categories (such as *Lingerie* or *Sexuality*), it is surprising that over 30% of *Social Media* sites and over 40% of dating sites were also labeled (and rejected) as *Pornography*. Another apparent issue is that domains related to *URL Shorteners*, *Video Sharing* or *File Storage* can often be related to other categories, such as *Music*, *Movies* or *Pornography*. This shows that deciding the correct label for a given domain can be hard, with the differences between categories being vague. Furthermore, not all users might behave honestly, as some could mislabel domains to pollute the system or gain advantages over competitors, *e.g.,* a pornographic site



(a) Adult content cluster

(b) Internet utilities cluster

Figure 7: Examples of overlap between categories in OpenDNS. The heatmap shows the frequency of X-axis categories being rejected when the Y-axis category is approved.

trying to be labeled as a video sharing site, or a company labeling a competitor's website as malicious or pornographic.

In the case of labels that often appear accepted together, we also find a high correlation among categories that could be related to sexual or nudity content (*e.g., Pornography, Nudity, Bikini/Lingerie*). Another interesting case is the pair *Advertising* and *Business Services*, which are accepted together over 30% of the time. This can be a result of many of these *Business services* acting as third parties offering advertising and tracking services too. Similarly, *News and Media* and *Television* often appear together since television stations often act as news outlets.

5.3 Is labeling domains a trivial process?

We perform an experiment using the authors of this study in order to gain a better idea of the aforementioned challenges behind OpenDNS' labeling process. One member of the research team manually selected 200 hostnames, including 50 for which OpenDNS and McAfee provide semantically equivalent labels; 50 for which they disagree; 50 from the top-1k domains in our normalized rank; and 50 unpopular sites. For ethical reasons, we discarded domains with labels that could be uncomfortable or harmful for our human labelers (*e.g.,* child pornography, nudity, violence, drugs, weapons, and malware-related ones). The remaining authors manually visited each website and labeled it using the OpenDNS taxonomy and definitions. Each domain was labeled by two authors, adding a third labeler when there was disagreement in the first stage.

Disagreement between two labelers is relatively high at 35.5% of domains, reaching 90.5% agreement between at least two reviewers when a third labeler is introduced. When the final results are compared to OpenDNS categories, we observe that our process could only achieve 71% accuracy; in 80.5% of the cases, at least one labeler reported the same category as OpenDNS. This experiment, while not representative, illustrates some of the challenges that arise when humans are involved in the process, even for experts in network measurements and cybersecurity. Disagreement is the result of subjective factors caused by different perceptions and sensitivities, but also by the inherent ambiguity of many of the categories forming the taxonomy and the dual nature of many websites, for instance, blogs offering political content [79] or tourism boards advertising casinos [80].

Takeaway: *We analyze OpenDNS's ecosystem of voters and editors, and find that most labels were submitted during the early stages of the project. We show that most users (95%) submit labels for only a few domains but that, in general, workload is evenly distributed among moderators. In the case of Curlie, we find that peer review may suffer from the low number of editors, but that categories are still being updated regularly. Furthermore, we find that labeling strategies involving humans are bound to generate disagreements. In OpenDNS, there are domains with 58 not approved labels. Moreover, the slight differences among labels generate clusters of related labels that often appear rejected together* (i.e., *Adult themes, Lingerie/Bikini, Pornography and Sexuality). We show that labeling is a non-trivial job by running a small-scale manual classification experiment, in which we only achieve 71% accuracy compared to OpenDNS and find that two labelers disagree on 35.5% of domains—highlighting the subjective nature of labeling.*

6 CASE STUDIES

In this paper, we have shown that researchers often rely on domain classification providers to either understand the type of domains that they observe in their study [31] (*i.e.,* to better characterize their results) or to gather a field-specific corpus of domains [25] (§ 2). Next to that, core applications of domain classification services are outside the academic circles. They are often used in technical solutions for content filtering and threat intelligence, for example in parental control apps [81] and school networks [14], which require accurate identification of specific types of domains.

Therefore, in this section we aim to understand whether choosing one domain classification service over another can yield different results when selecting target domains or when classifying domains specific to a given category. We analyze the usefulness and aptness of domain classification services for three types of domains that are often analyzed by the research community: (1) advertising and tracking services; (2) websites offering adult content (*i.e.,* pornography and gambling sites); and (3) domains that belong to a Content Delivery Network (CDN) and hosting providers. Our approach starts with obtaining available sanitized domain category sets to identify which domains belong to each one of these categories. Then, we analyze the coverage as well as the labels assigned to these domains by different classification services to identify potential errors and inconsistencies. While such specialized lists are more appropiate for choosing a pool of websites that belong to a given category, we have seen that it is still common for academic papers to rely on classification services for website selection or classification [23, 25, 32].

Advertising and tracking services. As ground truth, we take a list of manually sanitized domains indexed in EasyList [82] and

EasyPrivacy [83].[5] However, these lists allow blocking traffic at a full URL level.[6] To reduce bias in our case study, we opt to account only for domains that are fully blocked by these lists, regardless of the full URL path. After a manual sanitization process, we study the labels from different classification services for the resulting 24,825 advertisement and tracking-related domains and manually extract the resulting labels semantically related to advertising and tracking applications (*e.g., Web Marketing* or *Advertisement*). Table 5 (two leftmost columns) shows that none of these services are able to correctly label most domains as tracking or advertising. Forcepoint presents the highest accuracy, which is barely higher than 15%, at the cost of sacrificing coverage (51.6%). While McAfee and FortiGuard have a higher coverage, they classify fewer than 10% of the domains as trackers. Most of the errors arise from tracking- or advertising-specific subdomains. For instance, all providers classify `airpushmarketing.s3.amazonaws.com` and `tracking.eurosports.com` using labels related to *hosting/CDNs* and *news/media/sports*, respectively.

Identifying adult content. We rely on two resources to gather domains related to adult content [14]. First, we rely on a manually labeled and sanitized list of pornographic websites from Vallina *et al.* [25]. Additionally, we compose a list of gambling sites extracted from three government websites [87–89]. By combining these two sources, we compile a manually vetted list of 3,519 domains related to web services typically considered as "adult content". The results (Table 5, middle columns) show that 5 services do a good job at identifying and correctly labeling webpages that host adult content: OpenDNS, McAfee, FortiGuard, Forcepoint and Dr.Web. Yet, there are substantial differences across services. Alexa, Trend Micro and Websense do not provide a label for the majority of the websites analyzed. Therefore, this case study also demonstrates that the choice of one provider above another can have severe implications in the number of domains classified as adult content. We also examine which other labels are usually assigned to adult content domains, finding a high correlation with those related to video sharing and streaming media. These labels are, in most cases, technically correct but they do not allow to identify these domains as pornographic. We also see that some services assign labels that imply maliciousness of adult domains (*e.g., malicious*, *spam*, or *not recommended*).

CDN and hosting provider related domains. Content delivery networks (CDNs) remain the dominant means for serving popular content and represent Internet *infrastructure.* While most domain classification services (*e.g.,* McAfee and Fortiguard) contain labels referring to CDNs or hosting providers, the *content* classification is often mixed with an *infrastructure* classification. As an example, one service can classify a CDN-hosted site as *content delivery network* while another derives a label from the site's content (*e.g., news* or *personal blog*). In order to measure differences in the classification strategies of different services, we select those domains in our dataset that are related to CDNs and hosting services. To do so, we pattern match the CNAME record of all domains against more than 80 CDN signatures from WebPageTest [90]. In total, we obtain a corpus of 2,858 domains, for which we compare the coverage across domain classification services. Table 5 (rightmost columns) shows that only McAfee and FortiGuard provide a label for the majority of these domains. Both services classify these domains based on their function rather that on their content (*e.g., Internet Services*, *information technology*, and *content services*). For the other services, the coverage is so low that it is difficult to discover a trend in the labels. Yet, it is still possible to find examples of labels related to the actual content of webpages hosted on these services (*e.g., News*, *Adult content*, or *Business*) as well as to the type of service provided. None of these classification strategies are right or wrong, but the choice of service translates in differences in terms of coverage and labels for CDN and hosting provider related domains.

Table 5: Coverage for different types of domains.

Type	Ad/Tracking N=24,825		Adult content N=3,519		CDN N=2,858	
Label	Any	Related	Any	Related	Any	Related
OpenDNS	16.9%	3.9%	88.2%	88.0%	3.4%	0.2%
McAfee	70.8%	3.7%	99.1%	97.6%	98.2%	84.7%
Fortiguard	78.7%	7.7%	99.8%	98.8%	93.0%	81.7%
Alexa	2.8%	0.1%	1.1%	0.1%	0.1%	0.1%
BitDefender	32.0%	2.8%	83.5%	65.8%	27.0%	27.0%
Forcepoint	51.6%	15.1%	97.1%	94.9%	29.9%	3.1%
Dr.Web	9.0%	0.0%	92.5%	92.4%	0.3%	0.0%
Trend Micro	7.4%	0.9%	12.1%	11.8%	0.4%	0.0%
Websense	2.8%	1.0%	4.6%	4.5%	0.2%	0.1%

Takeaway: *For specialized use cases, the choice of one domain classification service over another can significantly impact the accuracy of academic studies and the effectiveness of solutions relying on them.*

7 DISCUSSION

In this section, we extract actionable insights from our empirical results, discuss best practices for using domain classification services, and propose various solutions as future work to overcome their limitations.

Dealing with insufficient accuracy. The key observations of our study are that *i)* coverage varies substantially between services (§ 4.2) and *ii)* the classification accuracy is marred by inconsistent taxonomies (§ 4.3) and low agreement among providers (§ 4.4). These inherent limitations set a high barrier for their effectiveness in real-world applications as well as their usage in research. For highly targeted use cases, general-purpose classification services may fall short. For example, as shown in our case studies (§ 6), the choice of service impacts the number of correctly identified adult domains. It may therefore be necessary to either search or develop curated and manually labeled domain-specific lists. Furthermore, end users and researchers should carefully consider the implications of errors. In applications like content filtering, errors can lead to inappropriately restricting access to legitimate resources ('overblocking') or, conversely, allowing access to undesirable resources ('underblocking') [91, 92]. For example, aggressive adult content filters could block sexual health information [93] or, as in the recent case of Cloudflare's DNS resolver, LGBTQIA+ sites [94]. In the academic domain, researchers can also take into account how important classification is to their studies, *e.g.,* using domain

[5] Both used by the anti-tracking solutions AdBlock and AdBlock Plus [84, 85].

[6] *e.g.,* they would not block the `bbc.co.uk` webpage, but they would block any URL from this domain which contains the `tracker.js` file [86].

categories to provide context for a minor result vs. generating the list of domains on which they base their whole study. There are a few documented cases in which authors preferred their own classification over those of commercial services due to concerns regarding their accuracy and coverage [9, 28, 29, 45, 46].

Dealing with biases. Coverage and accuracy suffer from selection and interpretive biases respectively. Service purpose determines which and how domains are classified: a filtering service may better cover and differentiate malicious domains, while marketing- or discovery-oriented services may provide a more fine-grained label for popular sites. How labels are sourced also introduces biases. For automated solutions, these stem from deficiencies in the training sets for machine learning algorithms. In a manual classification process, these are induced by maintainability challenges as well as human interpretation (§ 5). There are cases where using a domain classification service can produce sound results. Yet, researchers should gain a proper understanding of potential biases in their chosen services to assess the limitations of applying them in specific domains, *e.g.,* by consulting the documentation. To empirically gauge the coverage and accuracy of the used service specifically for their studied domains, researchers can additionally manually inspect random subsets to determine whether the labeling is of sufficient quality to make its usage appropriate.

Dealing with inconsistencies. When using domain classification services, results must be interpreted and reported with care, to avoid introducing errors due to inconsistencies. Domain classification services exhibit varying characteristics, *e.g.,* whether they provide multiple labels, label subdomains differently, or regularly update labels (§ 3). Moreover, they may behave unexpectedly, such as by deviating from their documented taxonomies (§ 4.3). Users should therefore verify the output of the services, *e.g.,* by analyzing aggregate statistics or a randomly selected sample. Furthermore, the specific applications of services affect their taxonomies. The granularity and exact meaning of a label (even if it is syntactically the same) thus largely differs between services and directly impacts the effectiveness of any application or the results of any study. Studies based on domain classification should thus examine the labeling taxonomy in detail and report the meaning of the selected labels to prevent wrong or incomplete conclusions.

Aggregation of multiple domain classifiers. Many websites are complex entities: it is hard to reduce them into a single label. Researchers might be tempted to overcome the limitations of individual domain classifiers—both in terms of coverage as well as label accuracy—by combining the output of multiple services in a single analysis pipeline. While this might be useful in some scenarios (*e.g.,* threat intelligence aggregators such as VirusTotal), we identify multiple challenges that rule out simplistic aggregation strategies:

(1) If the goal is to improve overall coverage, aggregating various classifiers might not necessarily achieve this purpose, as we showed in § 4.2. The choice of classifiers should be informed by the size of the intersecting set. In addition, we found coverage to vary greatly depending on factors such as domain popularity or freshness.

(2) Different classifiers might provide complementary perspectives on a domain's nature, but the aggregation of their labels can be difficult since they come from different taxonomies with radically different purposes. Simply taking the union of the outputs might unnecessarily increase the constellation of labels and increase redundancy, since two services might use semantically-equivalent labels to reflect the same purpose or abstract concept. This could be aggravated by services developing multilingual taxonomies. Reconciling multiple taxonomies coherently might be cumbersome and difficult to scale, particularly if it must be done semantically.

(3) Determining what is a discrepancy among classifiers and what is just a different perspective on the nature of a website could also be challenging. A site can simultaneously be labeled as *porn*, *streaming*, and *CDN* by three different providers. Understanding the focus, sensitivities, limitations, classification methods, and intended label usage of each classification service is an unavoidable step to properly contextualize and meaningfully aggregate their outputs.

7.1 Limitations and Future work

While we showed how domain classification services' characteristics can vary significantly and often tend to be unfavorable, we are unable to quantify the quality of individual services due to a lack of comprehensive ground truth. We therefore avoid putting forward specific guidance on which services end users as well as researchers should prefer. We provide directions for future work that would bring us closer to such an evaluation.

While we have been able to compare labels between services by analyzing their diversity, understanding the *semantic agreement* of these labels would require developing a new taxonomy to which all labels across all services need to be translated, similar to how AVClass [95] automatically annotates malware samples with one semantically-equivalent label generated from multiple antivirus labels. This translation could occur manually, which may be more accurate, but comes at a higher maintenance cost when taxonomies change or an additional service is to be integrated. Alternatively, this taxonomy development could be (partially) automated through methods such as label normalization, heuristics [96], determining strongly coupled label pairs between services, or a semantic interpretation of existing labels through natural language processing. Anecdotally, we explored the latter method, but it generated a high false positive rate (*e.g., web spam* and *web hosting* could be reported as equivalent).

Beyond case studies, we do not broadly evaluate label *correctness*: even if all services agree on a label, it might still be wrong. An independently developed classifier can serve as a more trustworthy source of labels, against which the labels from other services could be compared. In order to cope with the large scale of the Internet, such a classifier would need to rely on automated methods, based on those developed in the state of the art, such as topic modeling [97]. Potential sources of ground truth are human-developed directories such as DMOZ and Curlie (as used in previous work [4, 98–100]) or the categorization of pages on Wikipedia (*idem* [101]). While an automated model may not be able to achieve perfect accuracy, its methods and performance can be disclosed transparently, improving the soundness of research that depends on it and enabling unbiased evaluation.

These steps could result in a classification service that researchers can rely upon to retrieve category labels obtained through a well-documented process and embedded in a vetted taxonomy. Such a service could either translate the set of labels from existing third-party classification services into labels from a custom taxonomy, or output the labels from a custom independent classifier. We consider both challenges to be interesting avenues for future work.

8 RELATED WORK

Web classification. One direction of research concerns methodologies and tools to classify websites automatically [5–9, 102]. In this regard, Qi *et al.* [5] studied features and algorithmic approaches used for automatic website classification. Beyond textual features, another approach [6] proposes using the web site's visual content for classification. Closer to our study are recent efforts to understand VirusTotal. Here, Peng *et al.* [103] studied how it classifies "phishing" domains, as well as quantified the quality of the results, finding discrepancies between results from VirusTotal results and those provided by direct sources [7]. Despite these efforts, the question on how domain classification services that are in widespread use work and differ, and how their different approaches impact study results remains open—a question that we study in this paper.

Internet measurement research methodology. Recent work has critically analyzed data sets and tools that are regularly used in Internet measurement research, in terms of the soundness and representativeness of results stemming from their usage as well as their enabling of reproducible studies [104]. Moreover, these studies formulate recommendations for how researchers should use them or propose improved solutions. For the selection of a representative sample of the Web from popular domains, the regularly used Alexa top sites list was shown to be unstable and easily manipulable [22, 63, 64]. Le Pochat *et al.* proposed the Tranco list as an alternative [64]. For retrieving Web site contents through Web crawls, Ahmad *et al.* developed a framework to compare crawlers based on varying technologies, finding that the choice of crawler may significantly impact measurements [105]. Zeber *et al.* compared crawlers with each other and with human user traffic, and found results to vary over time as well as across platforms [106]. We provide a similar assessment of domain classification services, as they can equally impact the results of research studies.

9 CONCLUSIONS

In this paper, we empirically and comprehensively analyze 13 domain classification services in order to study their labeling strategy and performance. We find that their limitations and shortcomings heavily affect their suitability and applicability, both for practical solutions and for academic studies, as demonstrated through our case studies. Coverage varies greatly between services and is insufficient for many types of domains. The lack of a common taxonomy and labeling behavior prevents a fair comparison and combination of services. Meanwhile, services using human labeling suffer from potential disagreements. We conclude with recommendations on how these services should improve, as well as a discussion on how to limit their deficiencies when using them.

ACKNOWLEDGMENTS

The authors of this paper would like to thank the anonymous reviewers of this paper as well as our shepherd Paul Barford for their valuable feedback. We also thank Wouter Joosen, Lieven Desmet, Tom Van Goethem, and Tobias Lauinger. We also want to thank VirusTotal for allowing us to use their academic API. This project was supported by the European Union's Horizon 2020 Innovation Action program (grant Agreement No. 786741, SMOOTH Project), the Spanish grants DiscoEdge (TIN2017-88749-R) and ODIO (PID2019-111429RB-C21 and PID2019-111429RB-C22), and the Region of Madrid grants EdgeData-CM (P2018/TCS-4499) and CYNAMON-CM (P2018/TCS-4566), co-financed by European Structural Funds ESF and FEDER. This research is partially funded by the German Federal Ministry of Education and Research project AIDOS (16KIS0976) and by the Research Fund KU Leuven. Victor Le Pochat holds a PhD Fellowship of the Research Foundation - Flanders (11A3419N). Marius Paraschiv is co-funded by the EU Social Fund and Comunidad de Madrid's Youth Employment Initiative (PEJD-2018-POST/TIC-8811).

REFERENCES

[1] Yahoo! 2014. *Yahoo! Directory.* https://web.archive.org/web/20141122194515/https://dir.yahoo.com/

[2] Curlie Project Inc. 2020. *Curlie.* Retrieved 2020-05-08 from https://curlie.org/

[3] AOL Inc. 2017. *DMOZ - The Directory of the Web.* https://web.archive.org/web/20170314000301/http://www.dmoz.org/

[4] 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

[5] 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

[6] Daniel López-Sánchez, Angélica González Arrieta, and Juan M. Corchado. 2019. Visual content-based web page categorization with deep transfer learning and metric learning. *Neurocomputing* 338 (2019), 418–431. https://doi.org/10.1016/j.neucom.2018.08.086

[7] 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

[8] Chen-Huei Chou, Atish P. Sinha, and Huimin Zhao. 2010. Commercial Internet filters: Perils and opportunities. *Decision Support Systems* 48, 4 (2010), 521–530. https://doi.org/10.1016/j.dss.2009.11.002

[9] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. 2018. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *NDSS Symposium.*

[10] Bitdefender. 2020. *Bitdefender.* Retrieved 2020-04-25 from https://www.bitdefender.com/

[11] Symantec Corporation. 2020. *Symantec Sitereview.* Retrieved 2020-04-28 from https://sitereview.bluecoat.com/

[12] McAfee LLC. 2020. *Customer URL Ticketing System.* Retrieved 2020-04-24 from https://www.trustedsource.org/

[13] OpenDNS. [n.d.]. *OpenDNS Community: Domain Tagging.* Retrieved 2020-04-25 from https://community.opendns.com/domaintagging/

[14] Federal Communications Commission. [n.d.]. *Children's Internet Protection Act (CIPA).* https://www.fcc.gov/consumers/guides/childrens-internet-protection-act

[15] R. S. Rosenberg. 2001. Controlling Access to the Internet: The Role of Filtering. *Ethics and Information Technology* 3, 1 (March 2001), 35–54. https://doi.org/10.1023/A:1011431908368

[16] Monica T. Whitty. 2002. Should Filtering Software be utilised in the Workplace? Australian Employees' Attitudes towards Internet usage and Surveillance of the Internet in the Workplace. *Surveillance & Society* 2, 1 (Sept. 2002). https://doi.org/10.24908/ss.v2i1.3326

[17] 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

[18] Google Inc. 2020. *About contextual targeting.* Retrieved 2020-05-08 from https://support.google.com/google-ads/answer/2404186

[19] DNSFilter Inc. 2020. *Webshrinker.* Retrieved 2020-04-27 from https://www.webshrinker.com/

[20] Melissa Gallo. 2016. *Taxonomy: The Most Important Industry Initiative You've Probably Never Heard Of.* IAB. https://www.iab.com/news/taxonomy-important-industry-initiative-youve-probably-never-heard/

[21] IAB Tech Lab. 2017. *Taxonomy.* https://www.iab.com/guidelines/taxonomy/

[22] Walter Rweyemamu, Tobias Lauinger, Christo Wilson, William Robertson, and Engin Kirda. 2019. Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research. In *20th International Conference on Passive and Active Measurement*. 161–177. https://doi.org/10.1007/978-3-030-15986-3_11

[23] Philippe Skolka, Cristian-Alexandru Staicu, and Michael Pradel. 2019. Anything to Hide? Studying Minified and Obfuscated Code in the Web. In *Proceedings of the World Wide Web Conference (WWW)*.

[24] Muhammad Ikram, Rahat Masood, Gareth Tyson, Mohamed Ali Kaafar, Noha Loizon, and Roya Ensafi. 2019. The chain of implicit trust: An analysis of the web third-party resources loading. In *Proceedings of the World Wide Web Conference (WWW)*.

[25] 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 *Proceedings of the Internet Measurement Conference (IMC)*.

[26] Rebekah Houser, Zhou Li, Chase Cotton, and Haining Wang. 2019. An investigation on information leakage of DNS over TLS. In *Proceedings of the International Conference on Emerging Networking EXperiments and Technologies (CoNEXT)*.

[27] Ceren Budak. 2019. What happened? The Spread of Fake News Publisher Content During the 2016 US Presidential Election. In *Proceedings of the World Wide Web Conference (WWW)*.

[28] Jannick Sørensen and Sokol Kosta. 2019. Before and after GDPR: The changes in third party presence at public and private European websites. In *The World Wide Web Conference*.

[29] Emily Stark, Ryan Sleevi, Rijad Muminovic, Devon O'Brien, Eran Messeri, Adrienne Porter Felt, Brendan McMillion, and Parisa Tabriz. 2019. Does certificate transparency break the web? Measuring adoption and error rate. In *2019 IEEE Symposium on Security and Privacy (SP '19)*.

[30] Sergio Pastrana, Alice Hutchings, Daniel Thomas, and Juan Tapiador. 2019. Measuring eWhoring. In *Proceedings of the Internet Measurement Conference (IMC)*.

[31] Hang Hu, Peng Peng, and Gang Wang. 2019. Characterizing pixel tracking through the lens of disposable email services. In *IEEE Symposium on Security and Privacy (SP)*.

[32] Damilola Ibosiola, Ignacio Castro, Gianluca Stringhini, Steve Uhlig, and Gareth Tyson. 2019. Who watches the watchmen: Exploring complaints on the web. In *Proceedings of the World Wide Web Conference (WWW)*.

[33] Peng Peng, Limin Yang, Linhai Song, and Gang Wang. 2019. Opening the blackbox of virustotal: Analyzing online phishing scan engines. In *Proceedings of the Internet Measurement Conference (IMC)*.

[34] Hiroaki Suzuki, Daiki Chiba, Yoshiro Yoneya, Tatsuya Mori, and Shigeki Goto. 2019. ShamFinder: An Automated Framework for Detecting IDN Homographs. In *Proceedings of the Internet Measurement Conference (IMC)*.

[35] Xianghang Mi, Xuan Feng, Xiaojing Liao, Baojun Liu, XiaoFeng Wang, Feng Qian, Zhou Li, Sumayah Alrwais, Limin Sun, and Ying Liu. 2019. Resident evil: Understanding residential IP proxy as a dark service. In *IEEE Symposium on Security and Privacy (SP)*.

[36] Matthew Joslin, Neng Li, Shuang Hao, Minhui Xue, and Haojin Zhu. 2019. Measuring and Analyzing Search Engine Poisoning of Linguistic Collisions. In *IEEE Symposium on Security and Privacy (SP)*.

[37] Victor Le Pochat, Tom Van Goethem, and Wouter Joosen. 2019. Funny accents: Exploring genuine interest in internationalized domain names. In *Proceedings of the International Conference on Passive and Active Network Measurements (PAM)*.

[38] Gong Chen, Wei Meng, and John Copeland. 2019. Revisiting mobile advertising threats with MAdLife. In *Proceedings of the World Wide Web Conference (WWW)*.

[39] Francesco Marcantoni, Michalis Diamantaris, Sotiris Ioannidis, and Jason Polakis. 2019. A Large-scale Study on the Risks of the HTML5 WebAPI for Mobile Sensor-based Attacks. In *Proceedings of the World Wide Web Conference (WWW)*.

[40] Hsu-Chun Hsiao, Tiffany Hyun-Jin Kim, Yu-Ming Ku, Chun-Ming Chang, Hung-Fang Chen, Yu-Jen Chen, Chun-Wen Wang, and Wei Jeng. 2019. An Investigation of Cyber Autonomy on Government Websites. In *Proceedings of the World Wide Web Conference (WWW)*.

[41] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 2019. 4 years of EU cookie law: Results and lessons learned. In *Proceedings of the Privacy Enhancing Technologies Symposium (PETS)*.

[42] SurfControl. [n.d.]. *Juniper Test-a-Site*. Retrieved 2020-04-30 from http://mtas.surfcontrol.com/mtas/JuniperTest-a-Site.php

[43] Robert Mullins. 2007. *Websense makes $400M bid for SurfControl*. Computerworld. https://www.computerworld.com/article/2545021/websense-makes--400m-bid-for-surfcontrol.html

[44] Sazzadur Rahaman, Gang Wang, and Danfeng Yao. 2019. Security Certification in Payment Card Industry: Testbeds, Measurements, and Recommendations. In *Proceedings of the ACM Conference on Computer and Communication Security (CCS)*.

[45] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. In *26th Annual Network and Distributed System Security Symposium (NDSS '19)*. 15. https://doi.org/10.14722/ndss.2019.23378

[46] Charles Reis, Alexander Moshchuk, and Nasko Oskov. 2019. Site isolation: process separation for web sites within the browser. In *Proceedings of the USENIX Security Symposium*.

[47] Duncan Riley. 2016. *Websense acquires Stonesoft from Intel Security, renames combined company Forcepoint*. SiliconANGLE. https://siliconangle.com/2016/01/14/raytheon-websense-acquires-stonesoft-from-intel-security-renames-combined-company-forcepoint/

[48] OpenDNS. [n.d.]. *OpenDNS Community: Domain Tagging: Categories*. Retrieved 2020-04-25 from https://community.opendns.com/domaintagging/categories

[49] 2010. *TrustedSource Web Database, Reference Guide, Category Set 4*. Reference Guide. McAfee Inc. https://www.trustedsource.org/download/ts_wd_reference_guide.pdf

[50] Fortinet Inc. [n.d.]. *FortiGuard Labs: Web Filter Lookup*. Retrieved 2020-04-24 from https://fortiguard.com/webfilter

[51] Fortinet Inc. [n.d.]. *Web Filter Categories*. Retrieved 2020-04-24 from https://fortiguard.com/webfilter/categories

[52] Chronicle Security. [n.d.]. *VirusTotal*. Retrieved 2020-04-24 from http://www.virustotal.com/

[53] Alexa Internet Inc. 2020. *Alexa: Top sites by Category*. Retrieved 2020-04-24 from https://www.alexa.com/topsites/category

[54] Alexa Internet Inc. 2020. *Where do Alexa's Top Sites by Category come from?* Retrieved 2020-04-27 from https://support.alexa.com/hc/en-us/articles/200449844

[55] Bitdefender. [n.d.]. *Web Categories in GravityZone Content Control*. Retrieved 2020-04-25 from https://www.bitdefender.com/support/web-categories-in-gravityzone-content-control-2287.html

[56] Forcepoint Inc. 2020. *Real-time Threat Analysis with CSI: ACE Insight - Websense.com*. Retrieved 2020-04-25 from https://csi.forcepoint.com/

[57] Forcepoint. 2020. *Master Database URL Categories*. Retrieved 2020-04-25 from https://www.forcepoint.com/product/feature/master-database-url-categories

[58] Doctor Web. 2020. *Dr.Web*. Retrieved 2020-04-25 from https://www.drweb.com/

[59] Trend Micro Incorporated. 2019. *Trend Micro Site Safety Center*. Retrieved 2020-04-24 from https://global.sitesafety.trendmicro.com/

[60] Trend Micro Incorporated. 2019. *URL Filtering Categories for Worry Free Business Security Services (WFBS-SVC)*. Retrieved 2020-04-24 from https://success.trendmicro.com/solution/1059905-url-filtering-categories-for-worry-free-business-security-services-wfbs-svc

[61] Symantec Corporation. 2020. *Category Descriptions*. Retrieved 2020-04-28 from https://sitereview.bluecoat.com/#/category-descriptions

[62] Ken Carnesi. 2019. *APIs - Webshrinker*. Webshrinker. Retrieved 2020-04-27 from https://www.webshrinker.com/apis/#domain-category

[63] 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 *2018 Internet Measurement Conference (IMC '18)*. 478–493. https://doi.org/10.1145/3278532.3278574

[64] 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 *26th Annual Network and Distributed System Security Symposium (NDSS '19)*. 15. https://doi.org/10.14722/ndss.2019.23386

[65] Fortinet Inc. [n.d.]. *FortiGuard Web Filtering*. Retrieved 2020-04-24 from https://docs.fortinet.com/document/fortigate/6.0.0/handbook/120269/fortiguard-web-filtering

[66] Bitdefender. 2018. *Web Filtering Software Development Kit (SDK)*. Retrieved 2020-04-25 from https://download.bitdefender.com/resources/media/materials/2019/pan/en/Bitdefender-OEM-WebFiltering-SDK-Datasheet-creatent169-en_EN-Screen.pdf

[67] 2019. *Symantec WebPulse*. White Paper. Symantec. https://docs.broadcom.com/doc/webpulse-en

[68] 2019. *The Need for Threat Risk Levels in Secure Web Gateways*. White Paper. Symantec. https://docs.broadcom.com/doc/need-for-threat-risk-levels-in-secure-web-gateways-en

[69] DNSFilter Inc. 2019. *Behind the Curtain of DNSFilter's AI*. Retrieved 2020-04-28 from https://www.dnsfilter.com/wp-content/uploads/2019/04/How_Webshrinker_Works.pdf

[70] Google. 2020. *Chrome User Experience Report*. https://developers.google.com/web/tools/chrome-user-experience-report

[71] Common Crawl Foundation. 2020. *Common Crawl*. https://commoncrawl.org/

[72] 2020. *Domain Name Industry Brief*. Report Volume 17, Issue 1. Verisign. https://www.verisign.com/assets/domain-name-report-Q42019.pdf

[73] OpenDNS. 2011. *OpenDNS Community: Domain Tagging: Categories*. https://web.archive.org/web/20111108035057/https://community.opendns.com/domaintagging/categories

[74] Trend Micro. [n.d.]. *Website classifications*. Retrieved 2020-04-24 from http://solutionfile.trendmicro.com/solutionfile/Consumer/new-web-classification.html

[75] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory* (2nd ed.). Wiley-Interscience.

[76] Brian Hartvigsen. 2013. *Faster Domain Tagging And Approval*. OpenDNS Community - Idea Bank. Retrieved 2020-05-11 from https://support.opendns.com/hc/en-us/community/posts/220012547/comments/224509067

[77] AOL Inc. 2016. *DMOZ - Editorial Guidelines - Subcategories*. https://web.archive.org/web/20170303062930/http://www.dmoz.org/docs/en/guidelines/subcategories.html

[78] Curlie Project Inc. 2018. *Curlie - Editorial Guidelines - Subcategories*. Retrieved 2020-05-17 from https://curlie.org/docs/en/guidelines/subcategories.html

[79] politicalscienceforias2016. [n.d.]. *politicalscienceforias2016*. politicalscienceforias2016.wordpress.com

[80] Black Hawk Colorado Casinos. [n.d.]. *Homepage*. blackhawkcolorado.com

[81] Álvaro Feal, Paolo Calciati, Narseo Vallina-Rodriguez, Carmela Troncoso, and Alessandra Gorla. 2020. Angel or Devil? A Privacy Study of Mobile Parental Control Apps. *Proceedings of Privacy Enhancing Technologies (PoPETS)* 2020 (2020).

[82] EasyList. [n.d.]. *Overview*. https://easylist.to/

[83] EasyList. [n.d.]. *EasyPrivacy*. https://easylist.to/tag/easyprivacy.html

[84] AdBlock. [n.d.]. *Surf the web without pop ups and annoying ads!* https://getadblock.com/

[85] AdBlock Plus. [n.d.]. *The world's number 1 free ad blocker!* https://adblockplus.org/

[86] AdBlockPlus. [n.d.]. *Adblock Plus filters explained*. https://adblockplus.org/filter-cheatsheet

[87] Dirección General de Ordenación del Juego. [n.d.]. *Listado de URLs de operadores con licencia*. https://www.ordenacionjuego.es/es/url-operadores

[88] Belgian Gaming Commission. [n.d.]. *Official list of the Gaming Commission*. https://www.gamingcommission.be/opencms/opencms/jhksweb_en/establishments/Online/fplus/

[89] The official Isle of Man Government Site. [n.d.]. *Licence holders*. https://www.gov.im/categories/business-and-industries/gambling-and-e-gaming/licence-holders/

[90] WebPageTest. 2020. *optimization_checks.py*. https://github.com/WPO-Foundation/wptagent/blob/baab610/internal/optimization_checks.py#L62

[91] Marjorie Heins, Christina Cho, and Ariel Feldman. 2006. *Internet Filters: A Public Policy Report* (2nd ed.). Technical Report. Brennan Center for Justice. https://www.brennancenter.org/sites/default/files/2019-08/Report_Internet-Filters-2nd-edition.pdf

[92] Sarah Houghton-Jan. 2010. Internet Filtering. *Library Technology Reports* 46, 8 (Nov. 2010), 25–33.

[93] 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

[94] Matthew Prince. 2020. *The Mistake that Caused 1.1.1.3 to Block LGBTQIA+ Sites Today*. Cloudflare. https://blog.cloudflare.com/the-mistake-that-caused-1-1-1-3-to-block-lgbtqia-sites-today/

[95] Marcos Sebastián, Richard Rivera, Platon Kotzias, and Juan Caballero. 2016. AVclass: A Tool for Massive Malware Labeling. In *19th International Symposium on Research in Attacks, Intrusions, and Defenses (RAID '16)*. 230–253. https://doi.org/10.1007/978-3-319-45719-2_11

[96] 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

[97] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. 2019. Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. 149–166. https://doi.org/10.1145/3319535.3363200

[98] Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, and Eli Upfal. 1997. Web Search Using Automated Classification. Poster POS725. In *6th International World Wide Web Conference (WWW '97)*.

[99] Dunja Mladenić. 1998. Turning Yahoo into an Automatic Web-Page Classifier. In *13th European Conference on Artificial Intelligence (ECAI '98)*. 473–474.

[100] Hao Chen and Susan Dumais. 2000. Bringing Order to the Web: Automatically Categorizing Search Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00)*. 145–152. https://doi.org/10.1145/332040.332418

[101] Eda Baykan, Monika Henzinger, Ludmila Marian, and Ingmar Weber. 2011. A Comprehensive Study of Features and Algorithms for URL-Based Topic Classification. *ACM Transactions on the Web* 5, 3, Article 15 (July 2011), 29 pages. https://doi.org/10.1145/1993053.1993057

[102] Anton Akusok, Yoan Miche, Juha Karhunen, Kaj-Mikael Bjork, Rui Nian, and Amaury Lendasse. 2015. Arbitrary category classification of websites based on image content. *IEEE Computational Intelligence Magazine* 10, 2 (2015), 30–41.

[103] Peng Peng, Limin Yang, Linhai Song, and Gang Wang. 2019. Opening the blackbox of virustotal: Analyzing online phishing scan engines. In *Proceedings of the Internet Measurement Conference*. 478–485.

[104] Quirin Scheitle, Matthias Wählisch, Oliver Gasser, Thomas C. Schmidt, and Georg Carle. 2017. Towards an Ecosystem for Reproducible Research in Computer Networking. In *Reproducibility Workshop (Reproducibility '17)*. 5–8. https://doi.org/10.1145/3097766.3097768

[105] 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 *The Web Conference 2020 (WWW '20)*. 271–280. https://doi.org/10.1145/3366423.3380113

[106] David Zeber, Sarah Bird, Camila Oliveira, Walter Rudametkin, Ilana Segall, Fredrik Wollsén, and Martin Lopatka. 2020. The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing. In *The Web Conference 2020 (WWW '20)*. 167–178. https://doi.org/10.1145/3366423.3380104

[107] Andrea Morichetta, Martino Trevisan, and Luca Vassio. 2019. Characterizing web pornography consumption from passive measurements. In *Proceedings of the International Conference on Passive and Active Network Measurements (PAM)*.

[108] Santiago Vargas, Utkarsh Goel, Moritz Steiner, and Aruna Balasubramanian. 2019. Characterizing JSON Traffic Patterns on a CDN. In *Proceedings of the Internet Measurement Conference (IMC)*.

[109] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. 2019. Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers.. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.

[110] Adam Oest, Yeganeh Safaei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Kevin Tyers. 2019. Phishfarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists. In *IEEE Symposium on Security and Privacy (SP)*.

[111] Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2019. SciLens: evaluating the quality of scientific news articles using social media and scientific literature indicators. In *Proceedings of the World Wide Web Conference (WWW)*.

[112] Kang-Min Kim, Yeachan Kim, Jungho Lee, Ji-Min Lee, and SangKeun Lee. 2019. From small-scale to large-scale text classification. In *Proceedings of the World Wide Web Conference (WWW)*.

[113] OpenDNS. [n.d.]. *FAQ for Domain Tagging*. Retrieved 2020-04-25 from https://community.opendns.com/domaintagging/faq/

[114] Chronicle Security. [n.d.]. *VirusTotal - How it works*. Retrieved 2020-05-10 from https://support.virustotal.com/hc/en-us/articles/115002126889-How-it-works

[115] Alexa Internet Inc. 2015. *Is Popularity in the Top Sites by Category directory based on Traffic Rank?* Retrieved 2020-05-11 from https://support.alexa.com/hc/en-us/articles/200461970

[116] Forcepoint. 2020. *CSI: ACE Insight Frequently Asked Questions*. Retrieved 2020-04-25 from https://csi.forcepoint.com/Home/Faq

[117] Doctor Web. 2020. *Free Dr.Web online scanner for scanning suspicious files and links*. Retrieved 2020-04-25 from https://vms.drweb.com/online/

[118] Trend Micro. 2019. *Web Reputation Services (WRS) Lookup process in Officescan*. Retrieved 2020-04-24 from https://success.trendmicro.com/solution/1056324-web-reputation-services-wrs-lookup-process-in-officescan-osce

[119] Trend Micro. [n.d.]. *Website classifications*. https://web.archive.org/web/20111111013335/http://solutionfile.trendmicro.com/solutionfile/Consumer/new-web-classification.html

[120] Broadcom Inc. 2020. *[ALERT] 2019 Symantec Intelligence Services and WebFilter Category and Application Update*. https://knowledge.broadcom.com/external/article?articleId=185063

[121] Webshrinker. [n.d.]. *Website Category API Reference*. Retrieved 2020-04-27 from https://docs.webshrinker.com/v3/website-category-api.html

[122] Jason Koebler. 2017. *AOL Is Mysteriously Shutting Down the 19-Year-Old Community That Inspired Wikipedia*. Motherboard. https://www.vice.com/en_us/article/bmbd4m/aol-is-mysteriously-shutting-down-the-19-year-old-community-that-inspired-wikipedia

[123] AOL Inc. 2016. *DMOZ - Suggest a Site: FAQ*. https://web.archive.org/web/20170304122239/https://www.dmoz.org/docs/en/help/submit.html

A USAGE IN ACADEMIC STUDIES

Table 6: Usage of domains classification services in the literature in 2019. The "dependent" column indicates whether the results of the study depend on the quality of the service used.

Venue	Area	Papers	using service		# dependent	
			#	%	on service used	References
TMA	Measurements	24	0	0%	—	—
PAM	Measurements	20	3	7%	3	[22, 37, 107]
ACM IMC	Measurements	39	5	12%	5	[25, 30, 33, 34, 108]
NDSS	Security	90	1	1%	1	[109]
ACM CCS	Security	148	1	0.7%	1	[44]
USENIX Security	Security	112	1	0.9%	0	[46]
IEEE S&P	Security	90	4	4%	3	[31, 35, 36, 110]
PETS	Privacy	68	1	1%	1	[41]
ACM SIGCOMM	Networking / Systems	31	0	0%	—	—
ACM CoNEXT	Networking / Systems	32	1	3%	1	[26]
WWW	Web Tech.	360	9	3%	9	[23, 24, 27, 32, 38–40, 111, 112]
Total		**1,014**	**26**	**3%**	**24 (92%)**	

B PROVIDER ANALYSIS

We examine the claims made by classification services (if available) in terms of their purpose, methods used for classification, coverage of URLs and languages, and development of their taxonomy. We retrieve these details through a manual inspection of their own documentation.

OpenDNS. OpenDNS provides DNS-based content filtering, sourcing website categorization from its human volunteer-based "Domain Tagging" project [13]. Participants submit domains and their categories, on which other participants may vote; once the mapping of a domain to a category receives sufficient votes, it is available for approval by a community moderator before it is propagated to the content filtering system [76]. These moderators also review reports of incorrect categorization as well as categories of popular sites [113]. We expand on the effects of this voting procedure in § 5. OpenDNS has at least one confirmed category for almost 4 million domains, out of 12.7 million submitted domains [13]. A list of categories and short descriptions is available [48]. Users had the ability to suggest the addition of categories to the taxonomy [113]; it is unclear who approved these new categories.

McAfee. McAfee provides the "TrustedSource" online service (previously called "SmartFilter") for obtaining both the category and a reputation score-based risk assessment for a URL [12], mainly with the goal of client-side content filtering. A user of the service must choose one of eight 'products', which affects the 'URL Filter database' version used. Categories are specific to URLs. McAfee categorizes web pages through "various technologies", including both machine learning and manual review [49]. It is said to cover "millions of Internet sites" [49]. McAfee's category taxonomy is documented in detail, listing descriptions, examples and related categories as well as taxonomy updates [49]. However, this document was last updated in 2010.

FortiGuard. FortiGuard provides an online tool for retrieving content-based URL categorization [50], which supports the content filtering functionality in its FortiOS-based FortiGate firewall [65]. Websites are classified through a "combination of proprietary methods including text analysis, exploitation of the web structure, and human raters" [65]. FortiGuard's service is said to include over 45 million website ratings that cover over two billion URLs [65]. Categories are divided into seven high-level groups (adult, bandwidth-consuming, business, personal, potentially liable, security, and unrated), and short descriptions and test pages are available [51].

VirusTotal. VirusTotal is an online service providing analysis of potentially malicious files and URLs by aggregating the results from a large set of detection engines [52, 114]. It also lists the domain's category, but it is unique among the other services in that it does not establish its own categorization. Instead, it collects labels from existing services: at the time of our data collection, these were Alexa, Bitdefender, Dr.Web, Forcepoint, Trend Micro, and Websense, but since July 2020, these were (at least) Bitdefender, Comodo Valkyrie Verdict, Dr.Web, Forcepoint ThreatSeeker, Sophos, and Yandex Safebrowsing. For each service, VirusTotal displays at most one distinct label, without combining labels any further, *i.e.,* a domain can have as many categories as there are services. Categories are only provided for domains, even though a user can also request scanning for URLs.

Alexa. Alexa offers the ability to view the 500 most popular websites for a specific category [53], with a focus on marketing and content discovery. Its results are based on the human volunteer-based categorization from DMOZ [54], but in contrast to DMOZ, Alexa's lists only contain domains, not URLs. Alexa's taxonomy is also based on the DMOZ's taxonomy, but pruned to around 280,000 categories. Alexa does not allow searching for the category of a specific domain. The ranking within a category is calculated using the same methodology as the main Alexa top list, but if applicable only using the data for the specific subdomain [115]. As the main Alexa top list only lists base domains, this may result in a different relative ranking for two domains [115].

Bitdefender. Bitdefender provides content category-based website filtering in its consumer- and business-oriented products [10]. There is no free online categorization tool, but VirusTotal integrates Bitdefender's categorization into its domain analysis. Its database is said to cover millions of URLs in multiple languages [66]. A list of (ungrouped) categories, short descriptions, and examples is available [55].

Forcepoint/Websense. Forcepoint (renamed from Websense in 2016 [47]) provides an online tool for website threat and content analysis [56]. The tool shows both a static (*i.e.,* previously determined) and a real-time classification. The former results from a combination of automated and manual inspection [57], while the latter is based purely on an automated machine learning-based approach [56]. Forcepoint will classify the specific page of a given URL, not its base domain [116]. Categories are divided into six high-level groups (reputation, security, bandwidth-consuming, productivity-inhibiting, social networks, and baseline) for which short descriptions are available [57].

Dr.Web. Dr.Web includes a category-based website filter in its client-side anti-virus software, but its online tool only provides a binary classification of a URL's maliciousness [117]. A more detailed categorization is accessible through VirusTotal, but appears to only cover types of malicious behavior. No documentation is available on the categorization process or the possible categories.

Trend Micro. Trend Micro's classification security-oriented service is available online through its "Site Safety Center" [59]. Next to a content-based category, they establish a threat rating denoting whether a website is 'safe', 'dangerous', 'suspicious' or 'untested' [59]. Their database is said to include over 35 million URLs, and they acknowledge that "a few URL rating errors" may occur [118]. Trend Micro publishes two lists of available categories with short descriptions. One was last updated in late 2019 and appears to be used for their "Worry-Free Business Security" and "OfficeScan" web threat protection products [60]; its categories are grouped into seven 'filtering groups'. The other was published at the latest in November 2011 [119] and has not been updated since [74]; its categories are ungrouped.

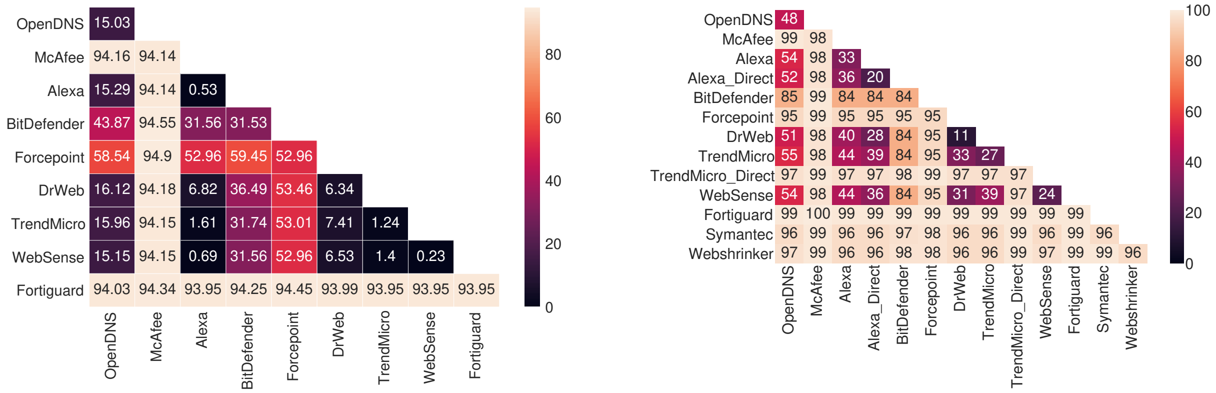
Symantec. Symantec (now part of Broadcom) provides an online tool to retrieve the URL categorization from its WebPulse system [11], which powers its web gateway content filtering. The categorization system is said to use manual and automated (machine learning) analysis, with several modules voting towards the final categorization [67, 68]. The tool indicates how recently the URL was categorized; previously unknown URLs are purported to be classified in real time [67]. Its URL database is said to cover "millions of entries", and supports over 60 languages [68]. A URL can be classified as up to four categories [67]. A listing of categories, descriptions, examples and test sites is available [61]. The taxonomy was last updated in August 2019 [120].

Webshrinker. Webshrinker provides an online demo tool of their URL categorization service [19]. Their service targets two audiences: a purely content-based categorization aimed at advertisers, and a security-oriented service which combines custom heuristics, machine learning, internal and external data feeds to assess web threats [19, 69]. Classification is said to occur in real-time [62], their database covers over 97.2 million 'entries' [69], and they support over 12 languages [62]. The two target audiences are also reflected in the two available taxonomies [62]. One is a custom list of 42 'standard' categories designed for content filtering, while the other uses the taxonomy of over 390 categories developed for marketing purposes by the Interactive Advertising Bureau (IAB) [21]. For the latter, Webshrinker computes a confidence score [121].

DMOZ. DMOZ (also known as the Open Directory Project) operated a directory of web pages, where users could navigate the category structure to find URLs in that category [3]. Its owner AOL took down DMOZ in 2017 after 19 years of operation [122]. DMOZ's rich taxonomy consisted of sixteen top-level categories, each being the top leaf in a large hierarchy of gradually more fine-grained subcategories, amounting to over a million categories encompassing 3.86 million URLs [3]. All users could suggest the addition of a URL to a category, but this had to be approved by one of the 91,929 category-specific editors [123]. Editors were also responsible for developing subcategories of the categories they maintained, which was suggested they do once a category reached 20 links [77]. DMOZ had strong multilingual support, with separate directories for 90 languages [3]. DMOZ allowed to search whether and where URLs appear in the directory.

Curlie. The Curlie project [2] emerged as the successor of DMOZ. Curlie retains the community of human editors, who appear to continue updating the directory listings and taxonomy to this day. Compared to its predecessor, Curlie has around 500 more categories (out of 1 million), but around 500,000 fewer URLs, and support for two more languages [2]. Like its predecessor, Curlie allowed to whether and where URLs appear search in the directory.

C LABEL UNION



(a) Union (4,424,142 domains)

(b) Union (Top-10K domains)

Figure 8: Coverage per service (diagonal) and the union of the coverage between pairs of services for our two domain sets (§ 4.1).

D RELATIONSHIPS BETWEEN PROVIDERS

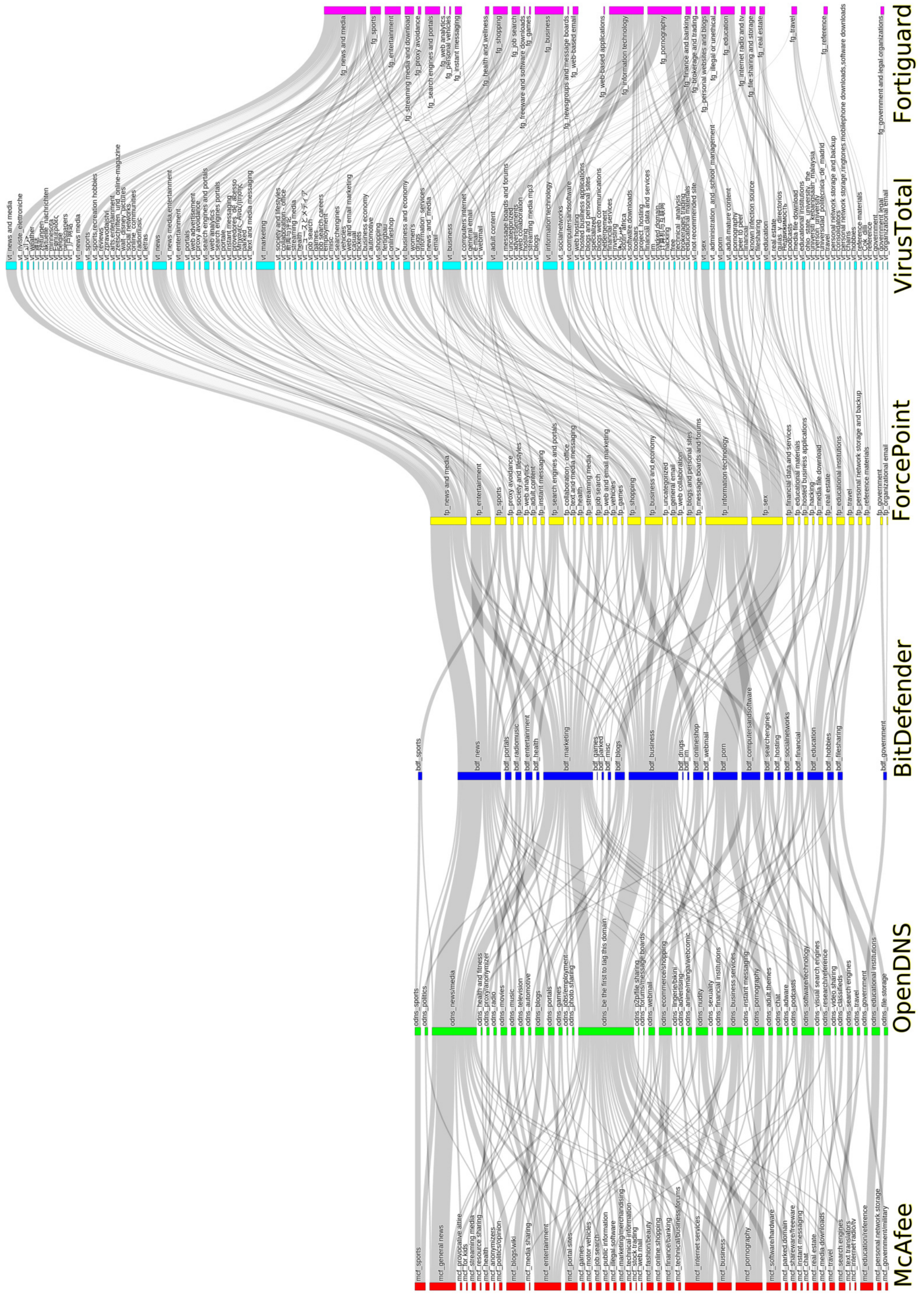


Figure 9: Label correspondences from top-1k domains for McAfee, OpenDNS, Bitdefender, Forcepoint, VirusTotal and FortiGuard.

E LABELS ACROSS SERVICES

When comparing pairs of services, it is instructive to also look at the cumulative distribution functions of one service over a corresponding one. These are presented in Figure 10. The horizontal axes contains all labels of a particular provider split into buckets, while the vertical axes represents the fraction of labels from the corresponding provider, covered by all the buckets up to the considered point. As expected, the curves for McAfee and OpenDNS (read row-wise) show a fast increase, as a small number of buckets contains the majority of labels, while Forcepoint and VirusTotal have a much more gradual increase. In some cases, a plateau appears at a point in the curve, as in the case of the Bitdefender-Forcepoint pair, or at the very beginning, as in the case of Bitdefender-McAfee. This is an artifact of the bucketing procedure which shows that the corresponding buckets cover a very small number of labels from the paired provider. This does, however, offer interesting information regarding labels that correspond on a one-to-one or one-to-few basis, even in the case of services that have a relatively reduced amount of overall labels.

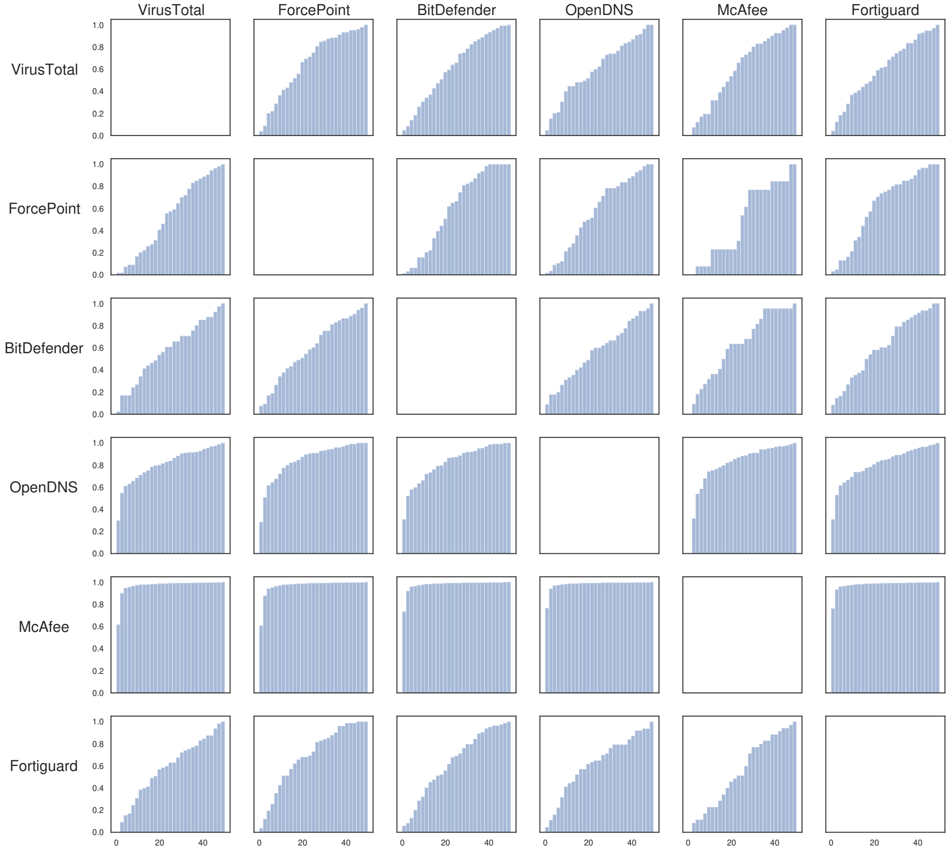


Figure 10: The distributions of labels for the six providers show considerable variation. Each row of the matrix represents the coverage of one provider in terms of the corresponding provider on the column. McAfee, Bitdefender and FortiGuard have a relatively small number of labels covering the set of domains, compared to the finer granularity of VirusTotal or Forcepoint. As to one label of McAfee, for example, there corresponds a considerable number of labels from VirusTotal, the conditional probability between pairs of labels from the two services is small, explaining the low values of conditional entropy as well as low mutual information. This is valid in all such one-to-many correspondences between providers.