Unmasking the Shadows: A Cross-Country Study of Online Tracking in Illegal Movie Streaming Services

Hussein Sheaib
Saarland University
h.j.sheaib@gmail.com

Anja Feldmann
Max Planck Institute for Informatics
anja@mpi-inf.mpg.de

Ha Dao
Max Planck Institute for Informatics
hadao@mpi-inf.mpg.de

ABSTRACT

The proliferation of Illegal Movie Streaming Services (IMSS) has posed significant challenges to legitimate streaming services and law enforcement alike, causing financial losses and complicating efforts to combat copyright infringement. Motivated by the absence of a comprehensive list of IMSS, and recognizing that IMSS websites often have short-lived domains, we first introduce a methodology to detect IMSS sites. Our evaluation demonstrates that our method achieves a recall of 84.31% in identifying IMSS. Applying this method on the Tranco Top 1M domains, we find 283 new websites hosting IMSS. When characterizing the IMSS ecosystem, our findings reveal that four specific IMSS sites attract considerable attention, appearing in the Tranco Top 10K domains. Additionally, these sites employ complex redirection patterns, with one site using up to 11 hops to evade detection. Using Google Identifiers, we then uncover 11 cases of co-ownership, where multiple sites share the same identifiers, indicating common operation. Finally, by crawling IMSS sites from seven vantage points (VPs), we investigate online tracking practices on these services — an area that has previously lacked thorough investigation. We find that more than 95% of IMSS include at least one third-party tracker on their websites. Interestingly, tracker presence is lower in the European Union (EU) countries than in other VPs. Furthermore, third-party tracking cookies are not the primary mechanism on IMSS sites; instead, the more invasive and unavoidable fingerprinting techniques are predominantly used for tracking across different VPs.

KEYWORDS

Measurement, web privacy, IMSS, Illegal Movie Streaming Services, IMSS ecosystem

1 INTRODUCTION

In the past few years, the online movie streaming business has bloomed. A Forbes report mentions that *99% of all US households pay for at least one or more streaming services* [18]. Netflix, for instance, the leading streaming service, has a total of 260.28 million subscribers worldwide according to the same report. This underscores the enormous market potential of the streaming industry, along with the significant revenue generation and job creation it contributes to any economy it operates within. As the market for legitimate streaming services expanded, movie piracy grew in parallel. Online movie piracy began in the early 2000s with peer-to-peer file-sharing services like BitTorrent [14]. Linking sites, which serve as directories of content hosted elsewhere, allow users to download files by clicking on links. Both types of services fall under the category of download-based piracy. The rise of Illegal Movie Streaming Services (IMSS) began in the early 2010s. According to a report by the Digital Citizens Alliance [3], IMSS sites accounted for 12.6% of all movie piracy sites in 2013. By 2014, this figure had increased to 17.8%, with revenue soaring from $27.5 million in 2013 to $46.2 million in 2014. By 2017, Muso [26] estimated that 80% of movie piracy was attributable to IMSS, demonstrating their significant exploitation of Internet infrastructure. These services use Internet infrastructure to operate covertly while attracting millions of users with seemingly free content. Additionally, users pay in non-monetary ways, as these services generate over $1 billion annually through advertising, data sales, and user tracking [1].

However, the enormous amount of visits these services receive [2] indicates that users often underestimate or deliberately disregard the substantial risks they face by accessing such services. Users can face security vulnerabilities due to poor hardware and software configurations [43], exposure to deceptive ads, malware, and scams in the free live streaming ecosystem [49], and the broader dangers of illicit online trade fueled by cyberlockers and darknet markets [11]. Additionally, frequent visits to infringing sites are strongly linked to increased malware downloads, with many users neglecting basic security measures [58]. Another significant aspect of IMSS sites that has not been fully explored is their use of invasive tracking mechanisms. While legitimate services are expected to comply with regulatory laws and provide privacy policies that offer transparency, IMSS operate outside these expectations and may aggressively track users without their awareness. Despite these growing concerns, little research has examined how tracking is conducted across different countries within the IMSS ecosystem.

This paper identifies sites that offer IMSS and investigate online tracking of their illegal ecosystem across seven different vantage points (VPs). The main contributions of the paper are as follows:

(1) We develop a lightweight method to automate IMSS site identification at scale by analyzing common features such as meta keywords, domain patterns, and DOM structures. Starting with a ground-truth dataset of 50 IMSS websites, we compile a list of 21 keywords and examine domain name variations across multiple Top-Level Domains (TLDs) to identify potential IMSS candidates. We implement an IMSS identifier that detects nine specific DOM patterns commonly used by these sites, allowing for accurate classification based on their layout and content organization. We evaluate our IMSS detection method, achieving 84.31% recall and identifying 282 new IMSS sites from the Tranco Top 1M domains. By combining the newly identified IMSS sites with those from

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–14
© YYYY Copyright held by the owner/author(s).
https://doi.org/XXXXXXX.XXXXXXX

the ground truth and evaluation datasets, our final dataset encompasses a total of 383 IMSS websites (§ 3).

(2) We characterize the IMSS ecosystem, including website ranking, TLDs, redirection patterns, hosting providers, and advertising practices. We find that 37% of IMSS sites are among the top 100K, with four sites ranked in the top 10K. These sites utilize a diverse range of 121 TLDs, with 31.3% employing at least one redirection before the user reaches the final destination. Additionally, *Cloudflare* dominates hosting IMSS sites, with 339 IMSS sites (88.5%) using it to host their websites. By examining real-time bidding digital advertising, we find that 35 IMSS sites (9.1%) serve valid *ads.txt* files. Among these, 77.1% declare *DIRECT* relationships with major ad networks, i.e., Google, as specified in their *ads.txt* files. Also, to reveal patterns of co-ownership and shared operation among IMSS sites, we analyze the Google publisher IDs and find that several sites share the same identifiers, suggesting they are operated by the same entities (§ 4).

(3) We investigate the online tracking ecosystem of IMSS sites among different countries. We find that more than 95% of IMSS on all VPs include at least one tracker. Interestingly, even though they are illegal services, we find that IMSS visited from the European Union (EU) countries link fewer trackers than from all other regions. Moreover, while third-party tracking cookies are present on fewer than 51% of IMSS sites, fingerprinting is much more prevalent, with over 70% of IMSS sites employing trackers that use this more invasive tracking method. IMSS sites prefer fingerprinting over cookies due to their resilience against user interventions, highlighting a distinctive reliance on this technique compared to the broader web, especially in countries like Australia, Brazil, and the United States (§ 5).

Finally, we address the risks users encounter when accessing IMSS sites, outline the limitations encountered during our research and discuss the ethical considerations of our study (§ 6).

2 RELATED WORK

2.1 IMSS ecosystem

The landscape of IMSS has evolved significantly since its inception on the Internet. Movie piracy existed long before online streaming services, with early methods including copying VHSs/DVDs and recording films in theaters [56]. With the advent of the Internet, piracy has become far more complex and widespread, now often involving organized crime rather than small-scale operations [59]. In 2017, Muso [26] revealed that more than 80% of online video piracy, comprising movie and TV shows piracy, was attributed to streaming piracy. This means that download-based options such as peer-to-peer networks were replaced by IMSS sites. Moreover, Blackburn et al. [8] found that movie piracy results in up to 560,000 jobs and up to $115.3 billion in reduced gross domestic product (GDP) losses yearly. In 2022, the Alliance for Creativity and Entertainment [2] reported 191.8 billion visits to movie and TV piracy websites, according to their published data on copyright infringement trends.

2.2 Digital piracy measurement

Several studies have focused on measuring and analyzing digital piracy on the Internet. Ibosiola et al. [29] concentrated their research on cyberlockers, uncovering a remarkably centralized system where a few networks, countries, and cyberlockers dominate content provisioning. Moreover, a report by the Digital Citizens Alliance [1] delved into the advertising ecosystem on movie piracy websites, revealing that Google's content delivery network accounts for 38% of all ads on piracy apps, with Google's ad services responsible for at least 13% of pirate ad placements, making it a significant contributor to revenue generation for these sites. Also, Ayers and Hsiao [28] investigated user tracking in illegal live streaming services, finding that tracking on these platforms is more pervasive and unavoidable than on legal streaming services, with deceptive ads and overlay redirects being common. Similarly, Rafique et al. [49] conducted a comprehensive analysis of free live-streaming services, also noting the prevalence of deceptive and malicious advertising on these platforms. Yang et al. [61] conducted the first in-depth analysis of illegal online gambling targeting Chinese users, uncovering the profit chain behind these activities. Their study revealed insights into the promotion strategies, payment methods, and infrastructure abuses that sustain this illicit ecosystem, providing valuable information for the security community to combat illegal online gambling. Moreover, Keshvadi and Williamson [32] investigated free live streaming services, which provide unauthorized broadcasts of live events, focusing on their behavior on Android smartphones. The authors analyzed 20 free live streaming services sports sites, examining packet-level data, video player performance, and privacy concerns, and compared them with legitimate online sports networks. Their findings revealed that free live streaming services often employ obscure tracking services, raising concerns about user privacy and service reliability.

There are substantial risks users might face by accessing such illegal services. Nikas et al. [43] highlighted the security risks faced by millions of users of illegal streaming services, primarily due to poor hardware and software configurations. It also introduced new attacks on these systems and discussed the forensic evidence that can be collected in such cases. Rafique et al. [49] identified the involved parties and their operational methods while mapping their impact on users in the free live streaming (FLIS) ecosystem. They revealed that users are exposed to deceptive ads, malware, and scams, while FLIS operators are frequently reported for copyright violations and often host their infrastructure in Europe and Belize. Chaudhry [11] highlighted the growing threat of illicit trade on the Internet, driven by cyberlockers and darknet markets, which has fostered a lucrative crimeware economy. Their work emphasized the need for multi-level enforcement, legislative updates, and private-sector initiatives to combat issues such as digital content piracy, ransomware, and malvertising. Telang [58] utilized a unique dataset from Carnegie Mellon's SBO project to analyze the causal relationship between frequent visits to infringing sites and increased malware downloads, finding that doubling time on such sites leads to a 20% rise in malware. Their study also revealed that users visiting infringing sites are less likely to take precautions, such as installing antivirus software, indicating a higher risk-taking behavior among these users.

2.3 Online tracking

Online tracking is the practice of collecting data about users' online activities. With the rise of technology and the Internet, data collection from individuals has become commonplace among advertising companies. Several web privacy studies have revealed the extensive nature and complexity of online tracking. Early research by Krishnamurthy and Wills [34] showed a significant increase in third-party tracking from 2005 to 2008. Subsequent studies documented the continuous growth and diversification of tracking techniques [38, 40, 48, 51, 52, 60]. Researchers have also explored various aspects of web tracking, including cookie syncing [48, 60] and fingerprinting techniques [7, 19, 44, 45]. These comprehensive studies have thoroughly documented the privacy risks associated with online tracking, raising concerns among individuals and privacy advocates. Previous research [6, 37, 55] indicates that people are increasingly worried about how companies collect and use their personal information. As a result, several data protection laws have been enacted to regulate the use of web cookies and other tracking and profiling techniques. These include the General Data Protection Regulation (GDPR) in the European Union (EU) [13], the California Consumer Privacy Act (CCPA) in California [10], and the General Personal Data Protection Act (LGPD) in Brazil [5].

Despite extensive research on online piracy, studies specifically focusing on IMSS remain limited. This work addresses this gap by developing a method for identifying IMSS sites and performing a detailed analysis of their ecosystem. Furthermore, while web tracking has been widely studied, there is a notable lack of research focused on tracking mechanisms within IMSS platforms. By conducting a rigorous examination of the tracking practices employed by these sites, this study provides new insights that contribute to both piracy research and the broader field of online tracking.

3 IMSS IDENTIFICATION

In this section, we present our method to identify IMSS websites (see Figure 1).

3.1 Ground-truth collection

We first collect a ground-truth dataset comprising 50 IMSS websites from the Piracy Megathread on Reddit[1] on 29 July 2024. This thread serves as a resource for individuals seeking information on various aspects of online piracy, specifically focusing on movies and TV shows. It consolidates links, tips, and guidelines related to streaming, downloading, and accessing pirated content.

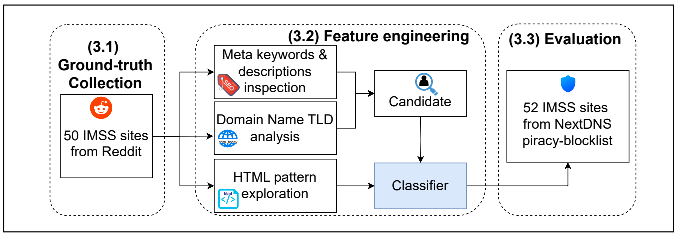


Figure 1: IMSS identification process.

#	Keywords
1	watch tv
2	free hd movies
3	watch hd movies online
4	watch movies
5	free movies
6	streaming anime
7	streaming movies
8	streaming tv
9	free films
10	watch free
11	free tv show
12	full movies free
13	watch latest movies
14	hd movies and tv shows, all free
15	download & watch online free
16	latest movies and tv shows online for free
17	stream free full movie online
18	tvshows online free
19	tv-series online for free
20	movies & tv shows for free
21	latest movies online for free

Table 1: IMSS-relevant keywords used by head crawler.

3.2 Feature engineering

Candidate identification: We then inspect the meta keywords and meta descriptions of each website and find that these websites use similar *head keywords*, which are popular keywords that drives high search volume. These keywords are crucial for maximizing website organic traffic while assisting the website in remaining competitive with others. For example, an IMSS website adds the tag <meta name="keywords" content="watch movies free"> to its head to get a higher rank in search engine results when users search for free movie websites. Following this inspection, we compile a list of general keywords by identifying frequently occurring phrases within the keywords meta tags of these sites. This approach allows us to capture commonly used terms across IMSS sites, creating a standardized keyword set for further analysis. Using this method, we finalize a list of 21 keywords to harvest potential IMSS candidates (see Table 1). Additionally, we incorporate the domain names of these well-known IMSS websites into the keyword list, as these famous sites are often indexed by search engines and may be used by other sites to enhance their visibility.

Furthermore, IMSS are inclined to use the same second-level domain (SLD) over several Top-Level Domains (TLDs). This practice involves registering the same base SLD with different TLDs to avoid disruption and evade legal measures. For example, if we identify *moviepirates.com* as an IMSS, we also consider similar domains such as *moviepirates.se* or *moviepirates.to* as potential IMSS since they all contain SLD *moviepirates*. Thus, any website that uses the same SLD in the ground truth is considered as a candidate for an IMSS.

[1] https://www.reddit.com/r/Piracy/wiki/megathread/movies_and_tv/#wiki_.1F4D1_.279C_streaming

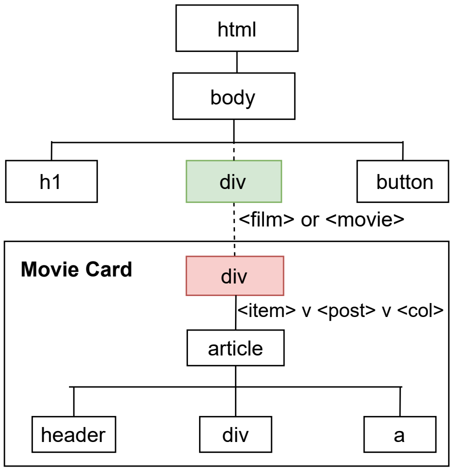


Figure 2: Example DOM pattern for IMSS sites.

Classifier building: Next, we observe that these IMSS websites use similar themes and layouts. These websites may be exact clones of each other with some minor tweaks, but in general, even IMSS with different domains still share several patterns. For this reason, we implement an automated script that serves the purpose of an IMSS identifier to search for nine common patterns on such websites on the DOM level. For example, Figure 2 shows a typical DOM pattern for an IMSS website. The pattern starts with the basic HTML and body tags, which contain all the visible content on the page. At the top level, there is a division *<div>* containing *<film>* or *<movie>* in its class for content related to films or movies. When the structure encounters this division, it continues with further divisions that contain one of *<item>*, *<post>*, or *<col>* in their class. If this specific structure is found on a website, we consider it to be an IMSS site, as this pattern is characteristic of how IMSS sites organize and display their content. We also identify eight other patterns that consist of similar components but vary across different IMSS websites. The detailed list of these patterns is provided in Appendix A.

3.3 Evaluation

To evaluate our method, we use the IMSS list from NextDNS's piracy blocklist for video streaming[2], which contains 1,085 URLs. We filter out URLs that do not return a 200 status code and those that trigger bot detection blocks, resulting in a set of 51 verified IMSS websites. Additionally, we incorporate the top 1K domains from the Tranco list generated on 29 July 2024 [36] into our evaluation. Table 2 shows the performance of our method in identifying IMSS sites. The confusion matrix shows that our method correctly identifies 43 IMSS sites while misclassifying eight as non-IMSS. No false positives are observed, and 1K non-IMSS sites are accurately labeled as such. Our primary objective is to ensure the classifier captures as many IMSS sites as possible without missing them. Thus, we focus on recall as the key performance metric. Our method achieves a recall of 84.31%, effectively minimizing missed IMSS sites.

Actual / Predicted	IMSS	Non-IMSS	Total
IMSS	43 (TP)	8 (FN)	51
Non-IMSS	0 (FP)	1,000 (TN)	1,000
Total	43	1,008	1,051

Table 2: Confusion Matrix evaluation.

3.4 Detection and validation on the Tranco Top 1M domains

Our IMSS dataset currently comprises 50 websites gathered from a piracy-focused subreddit and 51 websites from the NextDNS blocklist, resulting in a total of 101 IMSS sites. To further expand this dataset for comprehensive analysis, we apply our detection method to the Tranco Top 1M domains generated on 29 July 2024 [36], excluding the 55 sites already included in our existing IMSS set. The Tranco list is a research-focused domain ranking system that combines multiple ranking sources to create an average, aiming to minimize the impact of manipulation on domain rankings. We then deploy a web crawler over two days, running ten parallel processes on a server, to collect meta keywords and meta descriptions from websites. By analyzing the collected meta information alongside second-level domain, we identify 851 potential IMSS candidates and pass them to the classifier, which detects 345 IMSS sites (see § 3.2). To further validate the performance of this classifier, we manually visit each of these positive predictions to verify if it is an IMSS website. We look into the content (movies or TV shows that are known not to be streamed for free) and the existence of a streaming option. Websites that only offer a download option are not considered in our final IMSS list. It turns out 282 of them (81.7%) are true IMSS sites, thus further demonstrating the classifier's performance.

Key results: In total, we have 383 IMSS websites among which 282 are previously unknown and thus newly identified.

4 IMSS ECOSYSTEM

In this section, we explore the IMSS ecosystem by analyzing factors such as ranking distributions, Top-Level Domains (TLDs), redirection patterns, hosting providers, and advertising practices.

4.1 Ranking

We first show the empirical cumulative distribution function (ECDF) of IMSS sites' ranks according to the Tranco list in Figure 3. We see that approximately 83% of the IMSS websites rank within the top 1M on the Tranco list, indicating that most of these sites receive notable traffic. Moreover, about 37% of the IMSS sites are within the top 200,000, suggesting that a substantial portion attracts considerable attention. Notably, four IMSS sites rank within the top 10,000: *lookmovie2.to* (rank 5,210), *hdtoday.tv* (rank 7,194), *fmovies.co* (rank 7,942), and *sflix.to* (rank 7,980) — demonstrating their widespread reach. Despite ongoing regulatory and enforcement efforts, the continued popularity of these IMSS sites underscores the persistent challenge in limiting access to these illegal platforms.

[2] https://github.com/nextdns/piracy-blocklists/blob/master/streaming-video

4.2 Top-level domain

We then break down the number of IMSS sites by TLDs in Figure 4. Our dataset of 383 IMSS sites shows a remarkable diversity of TLDs, with 121 distinct TLDs represented. Similar to previous findings showing that older TLDs like *.com* tend to host more bad content than newer generic TLDs [33], we observe that 13.3% of IMSS sites use *.com*. In addition, 21.7% of second-level domains are registered across multiple TLDs, likely to capture search traffic and broaden user reach (see Table 5 in the Appendix B). For example, *lookmovie* operates under 10 TLDs, *putlocker* under nine, and *fmovies* under eight. This tactic complicates enforcement efforts, as targeting and taking down domains associated with a specific IMSS "brand" does not effectively disrupt access if the same second-level domain exists across various TLDs. As a result, even if one version of the domain is removed, alternative versions remain accessible, complicating efforts to curb these IMSS platforms.

4.3 Redirection patterns

Next, we show the frequency of URL redirection making IMSS sites available under more than one domain in Figure 5. We find that 120 IMSS sites (31.3%) employ at least one redirection before the user reaches the final destination. Among these, 77.5% use single-hop redirects. Notably, we identified a redirect chain involving 11 consecutive redirections[3]. These redirections fall into two categories: intra-domain (within the same TLD+1) and inter-domain (across different TLD+1). Intra-domain redirections, comprising 59.2% of cases, suggest efforts to keep users within related domains. Inter-domain redirections (40.8%) likely serve to obscure the final destination further. We hypothesize that IMSS operators use complex redirection patterns for multiple purposes. For example, avoiding domain-based blocking enables IMSS sites to bypass filters that block specific domains associated with illegal activity. By using multiple hops across different domains, IMSS sites reduce the likelihood that all domains in the chain will be flagged, allowing them to remain accessible even if some domains are blocklisted. This approach also serves as a protective mechanism—if one IMSS domain is taken down, it can redirect users to another, thus maintaining continuity despite enforcement efforts.

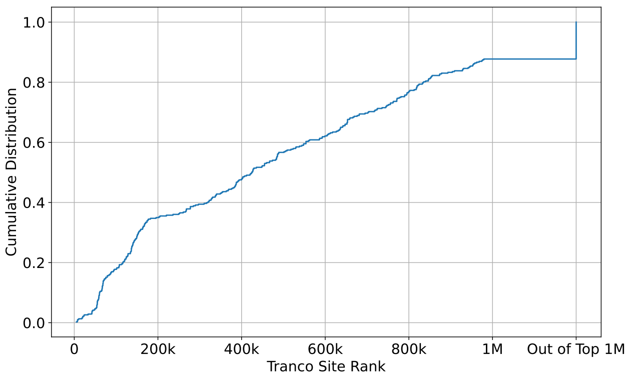


Figure 3: Empirical Cumulative Distribution Function of Tranco Rankings for IMSS websites (July 29, 2024).

4.4 Hosting company

Here, we investigate the companies behind the IP traffic of IMSS domain. To this end, we begin by capturing DNS requests for each IMSS domain to trace back to their respective IP addresses. With these IPs in hand, we proceed to identify the hosting providers by querying the Autonomous System (AS) numbers using the Ipinfo dataset[4]. Figure 6 shows the distribution of IMSS websites by the hosting company. Our findings reveal that *Cloudflare* is the most common host for IMSS, followed by *K4X* and *Amazon*, although these two host a significantly smaller number of sites.

We then manually review the Terms of Service (ToS) of these top providers. *Cloudflare* prohibits using their services for unlawful activities, but we found no specific filtering mechanisms to prevent IMSS sites from using their services. *Cloudflare* explicitly states that it is not a web host and has no direct control over the content displayed on websites using its services[5]. Similarly, *K4X*'s ToS prohibits illegal activities such as phishing, DDoS attacks, and malware distribution. *K4X* reserves the right to terminate services without prior notice for violations. However, we observed no specific filtering mechanisms targeting IMSS sites hosted by *K4X*. *Amazon*'s ToS also prohibits hosting illegal content and grants the company the right to remove or disable access to such content. As with the others, we observed no direct enforcement actions against IMSS sites during our investigation. Since the providers' ToS are similar, we hypothesize that one potential reason for *Cloudflare*'s dominance is its tier of free services[6], which may be particularly attractive to and exploited by IMSS.

Note that *Cloudflare* offers an abuse reporting mechanism[7], forwarding complaints to the website operator and hosting provider. *K4X* also provides an abuse reporting mechanism via email[8], allowing users to report malicious content. *Amazon* offers an Abuse Reporting Platform[9] for users to report violations, and they may take immediate action without prior notice in cases involving illegal content. However, we believe that hosting providers should apply filtering mechanisms to prevent their services from being exploited by IMSS websites and other illegal content.

4.5 Are IMSS sites reliable to the advertiser?

Here, we explore whether IMSS sites appear reliable to advertisers by examining one of the most common digital advertising methods used by online platforms: real-time bidding.

Real-time bidding provides the means for selling and buying advertising inventory on a per-impression basis. For this purpose, websites add **ads.txt** (Authorized Digital Sellers) to their root directory. This text file contains which entities are authorized to sell

[3] 123movies123.zone → 123movies123.today → 123movies123.movie → 123moviesn.com → to123movies.org → us123movies.com → 123moviesf.com → free-123movies.org → 123moviescc.com → 123movies.coach → movies123net.com → movies123pro.com

[4] https://ipinfo.io/

[5] https://community.cloudflare.com/t/piracy-site-hosted-by-cloudflare-needs-to-be-taken-down/248419

[6] https://www.cloudflare.com/plans/#

[7] https://abuse.cloudflare.com/

[8] https://phish.report/contacts/K4X

[9] https://support.aws.amazon.com/#/contacts/report-abuse

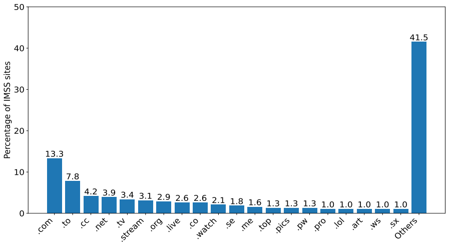


Figure 4: Distribution f IMSS sites by frequency of Top-Level Domains (TLDs).

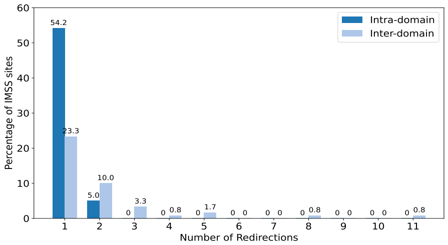


Figure 5: Distribution of IMSS sites by frequency of redirections.

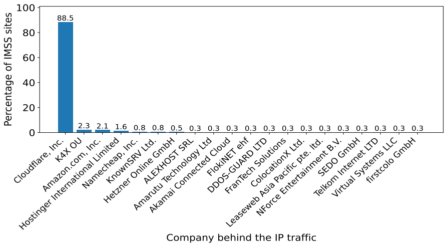


Figure 6: Distribution of IMSS sites by IP traffic company.

the publisher's, i.e., the website hosting the *ads.txt* file, digital ad inventory. Each entry in the *ads.txt* takes the following format:

Listing 1: Format of entries in *ads.txt*

```
<domain>, <publisher ID>, <relationship type>,
<certification authority ID>
```

The components of the entry are as follows:
<domain>: domain of the advertising entity (e.g., "google.com")
<publisher ID>: publisher ID or seller ID with the advertising entity
<relationship type>: either "DIRECT" or "RESELLER"
<certification authority ID>: Optional, indicates certification by a recognized authority

In addition to the *ads.txt* file, advertising entities also host a file called ***sellers.json***. This file lists the publishers selling their digital ad inventory. The use of both files—*ads.txt* and *sellers.json*—enhances transparency and accountability in the digital advertising ecosystem, helping to prevent ad fraud. These files are particularly important for *Supply-Side Platforms (SSP)* and *Demand-Side Platforms (DSP)*. SSPs are platforms that allow publishers to sell their digital ad inventory automatically, while DSPs enable advertisers to purchase digital ad inventory through a single interface.

To examine if the IMSS employs this method, we build a script to loop over the list of 383 different IMSS sites, append the link with /*ads.txt*, and store its file (if present). In total, we find 35 out of 383 IMSS sites (9.1%) serving a valid *ads.txt* file that follows the specification [35]. We parse and analyze the content of these files. Figure 7 shows the number of IMSS sites that hold direct and reseller relationships with a number of popular authorized digital sellers. These sites tend to form DIRECT business relationships with well-known ad networks, e.g., 77.1% of IMSS sites (27 out of 35) are hosting *ads.txt* with *google.com*. Although ad platforms are listed in these files, they may not always serve ads due to the dynamics of programmatic advertising. Nevertheless, a business relationship persists between the website and the ad network [47].

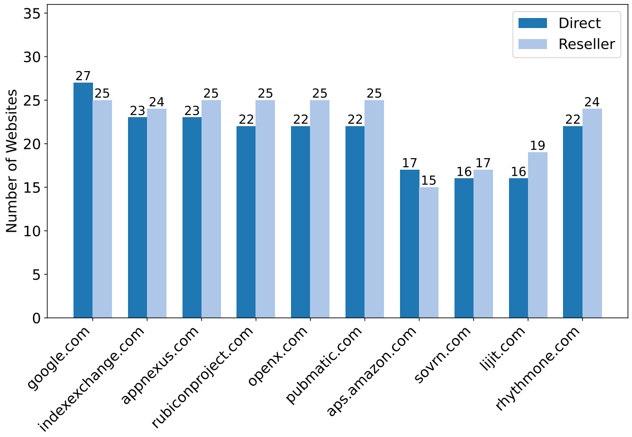


Figure 7: Authorized digital sellers of 35 IMSS sites hosting *ads.txt* in their root directory.

In addition, to verify the legitimacy of these relationships, we download the *sellers.json* file served by digital sellers (if present[10]) and inspect the IDs obtained from the DIRECT seller entries in *ads.txt*. The entries in *sellers.json* have the following format: {"seller_id": "", "seller_type": "", "domain": ""}. If the domain in *sellers.json* is the same as the domain hosting *ads.txt* for the same ID, this entry is verified. Out of a total of 188 direct seller entries, five entries (from two Google seller IDs) were verified using the aforementioned method, whereas all other sellers' relationships could not be verified. These two domains are *kuriname.com* and *sharmajazi.com*. It shows that two IMSS sites succeeded in creating a legitimate relationship with an ad leading company, i.e., Google, which shows that more filtering is required. Additionally, for the remaining unique direct 587 entries, 169 IDs are not found in the *sellers.json* and 418 IDs have a domain mismatch. This further highlights the need for improved verification and filtering mechanisms to ensure the integrity of digital advertising networks.

[10] List of digital sellers where we were able to locate and download the *sellers.json* file:

- Google: https://realtimebidding.google.com/sellers.json
- Amazon: https://aps.amazon.com/sellers.json
- AppNexus: http://acdn.adnxs.com/sellers/1d/appnexus/sellers.json
- Index Exchange: https://www.indexexchange.com/sellers.json
- Lijit: https://lijit.com/sellers.json
- OpenX: https://openx.com/sellers.json
- PubMatic: https://cdn.pubmatic.com/sellers/data/sellers.json
- RhythmOne: https://sellers.rhythmone.com/
- Rubicon Project: https://rubiconproject.com/sellers.json

Publisher ID	# IMSS sites	Site
UA-149357125-1 G-K10V17Z4DE	3	myflixer.today, myflixerz.cc, myflixer.cx
G-2TL7NH453R G-HJD8YWWX25	2	gomovies.pk, ullu.com.pk
UA-287393595-1 G-22128MS4HG	2	himovies.sx, himovies.to
G-PWBPFEY4VZ	2	upmovies.net, flixwave.me
G-FT0UYVTW9 G-FJ0KYLTM9 G-FL0MYNTO9 G-FV0WYXTY9	2	madstream.in, gomovies-hd.com
G-LB83DK5FX7	2	123moviesraw.co, 123moviesraw.com
G-S16ETKV	2	pelisflixhd.fun, pelisflixtv.lat
G-ZU5WL89ZK	2	pelisflix20.pro, pelisflix2.ong
G-CCAKM6A	2	pelisflixhd.fun, pelisflix.cfd
G-2EHJIFW	2	123moviesz0.com,fmovies0.cc
UA-1723892319-1	2	soap2daya.to,soap2dayto.ac

Table 3: Shared Google publisher IDs.

4.6 Co-ownership detection

Since the IMSS sites share the same HTML patterns, we hypothesize that they belong to the same administration. To detect IMSS website administration and co-ownership, we leverage the method based on publisher-specific IDs from Ref. [46].

We consider three Google publisher IDs, including the Google AdSense IDs that follow the format **pub-[0,9]{9,}**, Google Tag Manager IDs that follow the format **GTM-[A-Z0-9]{6,}**, and Google Analytics IDs that follow one of two formats: **G-[A-Z0-9]{7,}** known as measurement ID or **UA-[0-9]{,4}-[0-9]+** known as tracking ID. *AdSense* is a service intended for publishers to monetize their websites by displaying ads on their websites. Whereas *Google Tag Manager* (GTM) is a service that provides the means for analysts and developers to add and manage tags (JavaScript snippets or tracking pixels) on a website or mobile app without any need to modify code. Furthermore, *Google Analytics* is a service used by developers and website owners to track and measure website traffic.

We identify the Google Identifiers through an offline analysis of the collected data. Specifically, we use regular expressions to search for these identifiers within HTTP(S) requests and stored cookies. We find 368 unique pairs of IMSS sites and IDs of this sort through our IMSS list and per location across the different vantage points. We find 351 unique Google Analytic IDs (277 Measurement IDs and 74 Tracking IDs), eight unique Google Tag Manager Container IDs, and nine unique Google AdSense Publisher IDs. Furthermore, we search for different websites that share the same IDs. Table 3 shows 17 IDs that are shared among multiple IMSS sites. Nine of these IDs are shared among IMSS sites that possess a similar domain name, e.g., *myflixer.today*, *myflixerz.cc*, and *myflixer.cx* shared ID *UA-149357125-1*. The eight other IDs are shared over sites with different domain names. Four of these IDs are shared between *madstream.in* and *gomovies-hd.com*. Two IDs are shared between *ullu.com.pk* and *gomovies.pk*. One ID is shared between *upmovies.net* and *flixwave.me* and one between *123moviesz0.com* and *fmovies0.cc*. This indicates that the sites that share the same IDs are operated by the same entities. This approach allows us to identify only a small subset of the IMSS ecosystem, indicating that further analysis and more comprehensive methods are needed to capture the full scope.

Note that, we examined the WHOIS information of IMSS domains to gain additional insights about co-ownership; however, this effort yielded limited results, as privacy protection measures obscured registrant details across all IMSS sites.

5 INVESTIGATE ONLINE TRACKING OF IMSS SITES AMONG DIFFERENT COUNTRIES

5.1 Data collection

To characterize online tracking of IMSS sites, we use *OpenWPM* [21] to conduct automatic crawls on the 383 IMSS sites (see § 3.4). *OpenWPM* is a web privacy measurement framework that facilitates data collection for privacy studies. When users access a website, they often not only visit the main landing page but also navigate through other inner pages of the site. Additionally, some IMSS sites do not directly load their homepage to avoid being flagged by automated detection systems monitoring for illegal content. To fully capture the IMSS ecosystem, we customize the *OpenWPM* process to navigate to the homepage and subsequently proceed to an additional inner page (movie page). On the default page, we search for tags with an *href* attribute that suggests a redirection to the homepage, such as */home* or *home.html*. Additionally, the homepage button may contain text indicating the same, such as *Home Page* or *old* (referring to an older style of the website). If a homepage link is identified, we follow it; otherwise, we proceed to search directly for a movie page. In addition, identifying the correct movie page is challenging. Simply searching for a tag with an *href* attribute containing *movie* or *movies* might lead to pages listing all movies or trending movies, rather than the specific individual movie pages we seek. To address this, we employ two strategies: (1) we select three trending movies with unique titles and search for their corresponding href attributes; (2) we search for tags with an *href* attribute containing the word *movie* but ensured that at least two additional characters followed *movie*. We find that this approach reduces the likelihood of landing on an unwanted inner page (movie page).

We then use AWS cloud instances with two VPs inside GDPR countries (Germany and Sweden), one VP in the United States (which is under CCPA), and one in Brazil (which is under LGPD). We also set up one VP in Africa, one in Australia, and one in Asia for more diversity of geolocations. Since measurement results are not stable, querying a website twice usually yields two different metric values. According to [41], to have a relative standard error smaller than 1%, we visit each website 10 times in each mode and record all information. Typically, a crawl on the target IMSS sites takes about 12 to 14 hours with commodity hardware. We perform the crawls in August 2024 using default settings. All measurements collect HTTP requests and responses, cookies, and JavaScript calls. The crawls were executed in headless mode, where the browser runs without a graphical user interface, and in stateless mode, where no session data or cookies were preserved between page loads.

We note that our manual examination of IMSS sites found three sites displaying cookie banners. This finding significantly reduces the likelihood that a lack of interaction with these banners has substantially affected our analysis of cookies-based tracking.

5.2 Tracking resource detection

To categorize first-party and third-party resources, we utilize the public suffix list [42] based on the visited domain. Then, we analyze individual third-party requests using established advertising and tracking blocklists to identify requests belonging to tracking or advertising companies. These blocklists are designed to target advertising and tracking resources and are commonly used by browser extensions that aim to safeguard user privacy while browsing. We use three popular blocklists for identifying advertising and trackers from web pages: *EasyList [23]*, *EasyPrivacy [24]* and *adservers [39]* lists. We use the *Adblockparser* [53], which works with *Adblock Plus* filter rules, to match URLs against these lists to identify tracking resources. To determine if an extension would have blocked a request using these lists, we directly match the blocklist rules quoted above with all requests over ten runs among different VPs. We inspect individual third-party URLs using these blocklists from July 27th, 2024.

5.3 Online tracking analysis

5.3.1 Overview. As online tracking is a common technique on the Internet, even on legitimate websites, it becomes even more intriguing to explore how such tracking is employed on IMSS websites. We find that 95.58% of IMSS over all locations include at least one third-party tracking request. The summary of the measurement instance configurations is in Table 4. This number indicates the pervasive use of tracking techniques on IMSS, highlighting the dangers and privacy risks a user can face by accessing such platforms.

Figure 8 shows the distribution of the average number of total HTTP requests (blue boxplot), number of third-party requests (orange boxplot), and number of third-party tracking requests (green boxplot) per website across seven VPs. Websites that have less than five total requests are discarded as it is assumed that either the website was not reachable in that visit or that the interaction attempt to leave the default page (navigating between the default, home, and movie pages) was not successful. We first observe that Germany and Sweden exhibit the lowest median in the number of requests across all three metrics, suggesting that IMSS websites in these countries generate fewer requests on average. This trend may be influenced by the GDPR in the EU, which could discourage extensive tracking and request generation. In contrast, the United States shows the highest median, reflecting that IMSS websites in this VP tend to generate a significantly higher volume of requests. Furthermore, the distribution of total requests (blue boxplot) shows a positive skew in most countries except Germany and Sweden. In these two countries, the median number of requests is centered within the boxplot, indicating a more balanced distribution of request volumes with fewer extreme values. In other locations, particularly in the United States, India, and Brazil, the positive skew suggests that while most IMSS sites generate a moderate number of requests, a smaller subset generates a much larger volume. This pattern suggests the presence of sites with more intricate or extensive operational setups in these regions. Regarding third-party requests (orange boxplot) and third-party tracking requests (green boxplot), the medians across most locations are either at or below the center of the boxplots, indicating a more balanced distribution compared to total requests (blue boxplot). Finally, we observe that the outliers are present across several locations, suggesting that certain IMSS websites consistently generate a significantly higher number of requests than others within the same country. It may represent sites with more complex structures, multiple third-party integrations, or more aggressive tracking mechanisms.

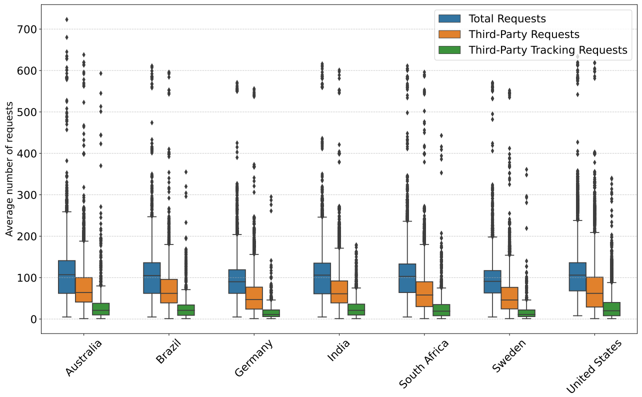


Figure 8: Distribution of total requests, third-party requests, and third-party tracking requests per IMSS website over ten runs and across different vantage points.

Websites may incorporate multiple third-party services and thus may include multiple trackers in their websites. Therefore, we show the distribution of third-party tracking domains per website among different locations and across the 10 runs in Figure 9. In general, across all countries, the distribution is heavily skewed toward the left side of the x-axis, indicating that most IMSS sites are associated with a relatively small number of unique third-party tracking domains. Specifically, the majority of IMSS sites are linked to fewer than 20 unique trackers, suggesting that these sites tend to use specific trackers tailored to their purposes. In Australia, approximately 17% of IMSS sites embed three distinct trackers—the highest observed percentage for that country. Brazil shows an even higher rate, with around 21% of IMSS sites embedding three trackers. India, South Africa, and the United States also have their highest percentages associated with three-tracker embeddings: 18%, 17.4%, and 17% respectively. Germany and Sweden, however, exhibit a different pattern. In these countries, the highest percentage of IMSS sites embed only one tracker—nearly 16% in Germany and approximately 17.5% in Sweden. This variation might stem from differences in regulatory environments or site operational practices. Furthermore, we find one website in the United States that included 51 different trackers, with similar cases in Brazil and South Africa, where a site embedded 51 and 50 trackers, respectively. This again shows a significant difference between the two EU VPs and the others, where most IMSS sites have five or fewer unique trackers. Whereas in the other VPs, at least 16% of IMSS sites serve requests from 10 or more tracking domains. We hypothesize that despite being illegal, IMSS operators may still reduce the number of trackers in the EU due to several factors. First, they may aim to circumvent

#	Vantage Point	# Total Reqs	# Third-party Reqs	# Third-party Tracking Reqs	#Third-party Domains	#Third-party Tracker Domains
1	Australia	147,848	116,309	66,637	877	462
2	Brazil	133,184	102,292	54,113	835	431
3	Germany	115,435	80,640	35,428	765	375
4	India	136,967	106,244	57,267	811	414
5	South Africa	146,622	110,643	60,754	861	463
6	Sweden	115,619	79,487	33,573	765	378
7	United States	167,177	132,055	82,083	853	462

Table 4: Census measurement configurations. The measurements were run concurrently on different AWS. The data was calculated by the unique metric values over 10 runs.

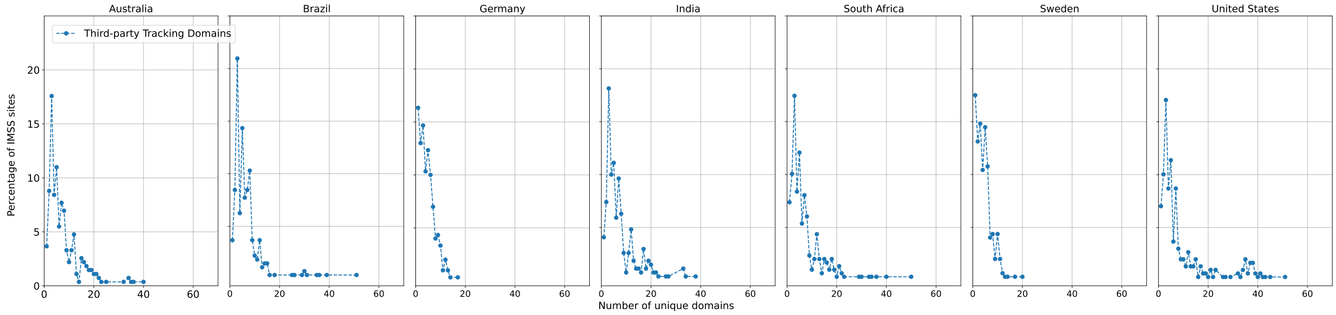


Figure 9: Number of third-party tracking domains per website among different vantage points. To get the stable metrics, the data was calculated over 10 runs.

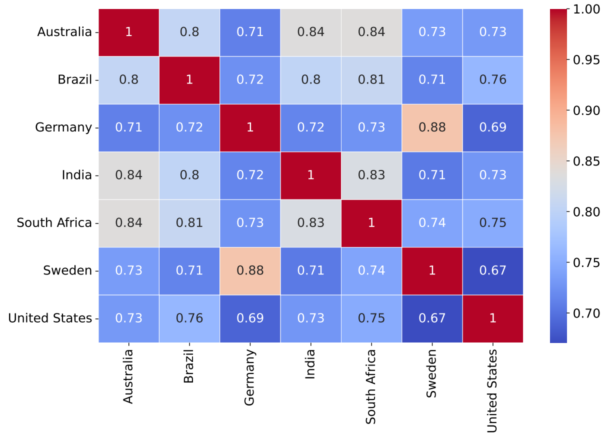


Figure 10: Jaccard similarity between tracker domains across different vantage points. The heatmap shows the degree of overlap in tracker domains between the VPs, with values closer to 1 indicating higher similarity and values closer to 0 representing lower overlap.

detection and legal action by minimizing exposure to strict regulations like the GDPR. In addition, many IMSS rely on third-party advertising networks that may impose compliance requirements with EU regulations, indirectly influencing the tracking practices of these sites. This aligns with previous findings on how geographic location influences tracking cookies [22, 50].

5.3.2 *Trackers on IMSS sites.*

Here, to evaluate the similarity of tracking behaviors across different geographic regions, we collected tracker domains from seven VPs. Using Jaccard similarity [30], we compare the overlap of tracker domains across these VPs to assess the consistency and divergence of tracking practices in Figure 10. The Jaccard similarity metric measures the proportion of shared tracker domains between any two VPs relative to the total number of unique domains across both VPs. We observe that Germany and Sweden have a high Jaccard similarity score of 0.88, indicating a significant overlap in the tracker domains used in these two VPs, which are in GDPR-regulated regions. In contrast, the United States exhibits lower similarity scores with most other VPs, such as 0.69 with Germany and 0.67 with Sweden, reflecting a more distinct set of tracker domains and potentially divergent tracking practices.

We then show the top 10 most popular third-party tracking domains/organizations and their prevalence among different locations over 10 runs in Figure 11. We show that Google-related domains such as *googletagmanager.com* and *google-analytics.com* are consistently among the most prevalent trackers across VPs. Compared to the top tracking domains and organizations identified on popular websites [15, 25], only Google is consistently present among the top tracking domains on IMSS sites across seven VPs. In the United States, the overlap extends to include other players such as Microsoft and Oracle. Notably, the leading companies such as Facebook and Twitter are absent from IMSS sites, highlighting that the tracking ecosystem on IMSS sites differs significantly from that of the broader web. Moreover, in all VPs except Germany and Sweden, five trackers are present on more than 40% of IMSS sites. Whereas in Germany and Sweden, only three trackers are present on more

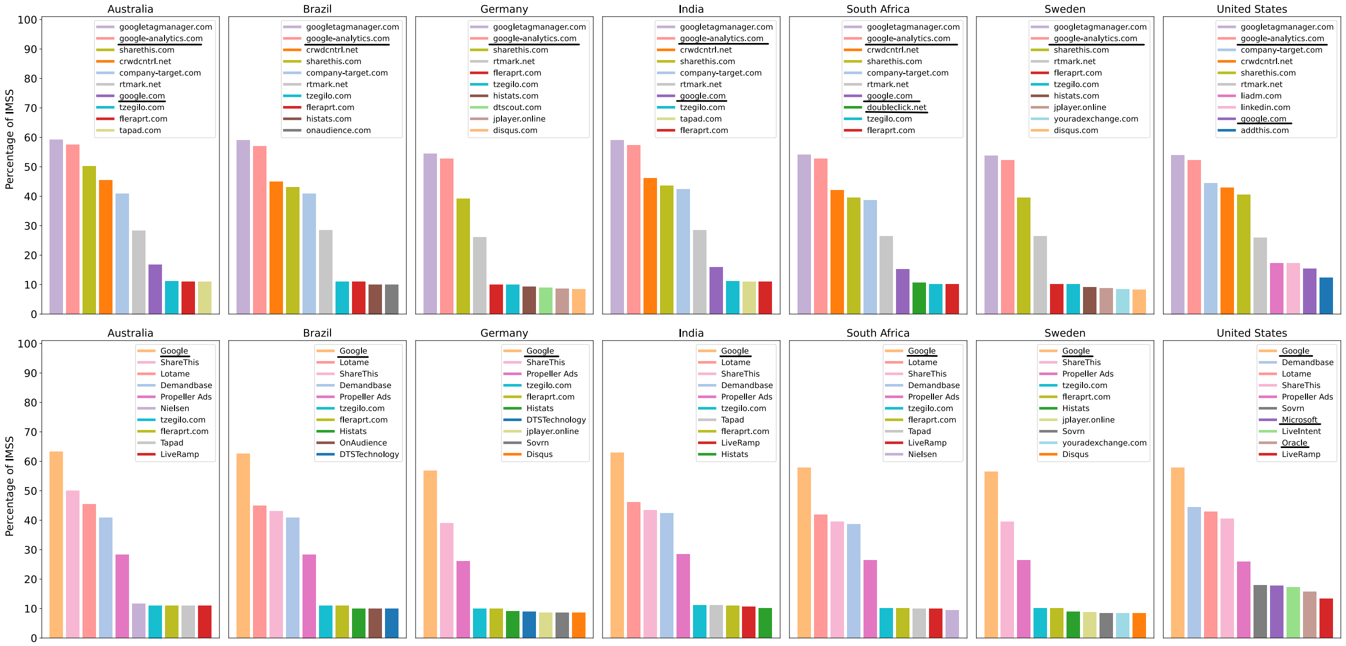


Figure 11: Top 10 most popular third-party tracking domains/organizations and their prevalence across different VPs. Domains/organizations that are underlined in black also appear among the top domains/organizations for popular websites, as reported by [15, 25].

than 40% of IMSS sites. Note that, *linkedin* tracking requests are spread over around 20% of IMSS in the United States.

We also search for tracker domains that only appear in one location exclusively. They include both localized domains (e.g., google.com.au in Australia) from a global company and more obscure ones. The United States has the highest number of unique tracker domains, with 38 entries. This extensive list suggests a highly fragmented and competitive digital advertising landscape, where numerous entities are vying for user data. The variety of domains includes well-known platforms like *akamaihd.net* and niche services like *geniusmonkey.com*, indicating a broad spectrum of tracking activities ranging from content delivery to behavioral analysis and ad targeting. Whereas there are only three unique domains in Germany and four in Sweden. The regional specificity of these tracker domains indicates a deliberate strategy by advertisers and data companies to tailor their activities to the unique characteristics of each market. This could involve adapting to local regulations, aligning with regional consumer behavior, or optimizing content delivery and ad targeting to maximize effectiveness. The list of unique tracker domains can be found in the Appendix C.

5.3.3 Are third-party tracking cookies the primary method used for tracking on IMSS sites? Since third-party cookies are the most currently known method to identify a user, we explore whether cookies serve as the primary mechanism for tracking on IMSS sites. To identify third-party tracking cookies, following the approach from Ref. [27, 50], we utilize the *justdomains* blocklist [31] to identify tracking cookies. This list, updated in August 2024, comprises of entries from popular tracking lists as in § 5.2. If a cookie's host matches any domain in the *justdomains* list, it is classified as a tracking cookie. In addition, we explore the usage of fingerprints on IMSS sites by leveraging the *Disconnect* tracker lists [17] and the collection of JavaScript APIs of potential fingerprinting. The fingerprinting techniques we detect rely on fingerprinting collection JavaScript APIs, including the *browser* [19], *HTMLCanvas* [4], *offlineAudio* and *WebRTC* [21], *screen* and *storage* information [9]. By analyzing these actual JavaScript calls, we ensure that our findings reflect active fingerprinting techniques, rather than merely flagging general tracking behavior. As a result, we have not overestimated the use of fingerprinting, as the captured data confirms whether the domains are actively engaging in fingerprinting practices. This approach provides a more accurate representation of the prevalence of fingerprinting on the studied IMSS websites.

We then show the third-party cookies and fingerprinting presence on IMSS sites in Figure 12. We first observe that third-party tracking cookies were found on less than 51% of IMSS sites across all locations, suggesting they might not be the primary tracking mechanism, likely because they can be easily deleted by users in the browser. In addition, again, Germany and Sweden show the lowest usage of tracking cookies, aligning with these regions' stringent data protection regulations like GDPR, which may limit the effectiveness or legal permissibility of cookie-based tracking. In contrast, fingerprinting is far more prevalent, with over 70% of IMSS sites employing trackers that use this method across all analyzed countries. Australia and Brazil show particularly high rates, with nearly 90% of IMSS sites serving trackers that use fingerprinting. However,

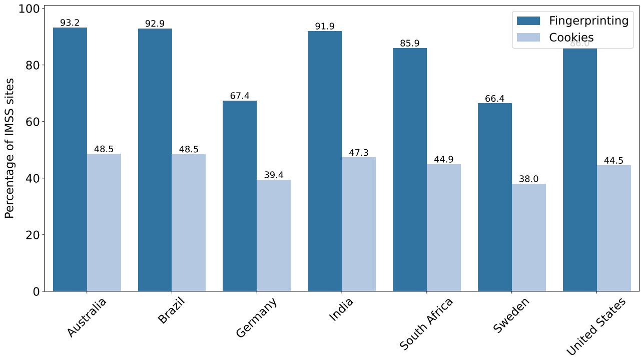


Figure 12: Prevalence of third-party cookies and fingerprinting on IMSS sites across different VPs.

according to the HTTP Archive report [11], the overall percentage of websites that use an external library to fingerprint their users is quite small. This highlights the distinctive nature of IMSS sites in their reliance on fingerprinting compared to the broader web. In addition, country-specific trends also reveal that IMSS sites in Australia, Brazil, and the United States heavily rely on fingerprinting. Germany and Sweden, despite slightly lower rates around 70%, still show substantial dependence on this technique. These differences may stem from regional regulations or variations in user behavior and browser settings. We can conclude that trackers may prefer fingerprinting mechanisms over cookies because these methods are more resilient to user interventions. While users can mitigate cookie tracking by clearing their browser data between sessions, fingerprinting is not as easily countered, making it a more reliable option for persistent tracking on IMSS sites.

To better understand the fingerprinting mechanism on IMSS sites, we investigate the fingerprinting techniques used by trackers. We find six fingerprinting techniques used by trackers on IMSS sites: *Canvas API*, *WebRTC API*, *OfflineAudioContext API*, *Navigator API*, *Screen API*, and *Storage API*. Obviously, IMSS sites utilize a wide range of established fingerprinting methods to track users. The detailed prevalence of the fingerprinting techniques on IMSS sites across seven locations are in the Appendix E.

Note that, we further investigate the use of the Topics API, a mechanism enabling interest-based advertising[12] across 383 IMSS sites and confirm its absence within our dataset.

6 DISCUSSION

6.1 The price of *free* on IMSS

Many studies examine the privacy and security risks of accessing illegal streaming sites [11, 43, 49, 58]. In this paper, we also pointed out that tracking is primarily accomplished through more persistent techniques like fingerprinting on IMSS sites. Therefore, our study exposes the privacy risks associated with accessing IMSS sites. Even in regions with stricter privacy regulations like the EU, where third-party tracking requests are less frequent, fingerprinting is present on approximately 70% of IMSS sites. This shows the magnitude of exposure users face by accessing these sites.

6.2 Limitations

Firstly, while our accuracy is promising, IMSS websites tend to evolve over time to evade detection. Frequent changes in domain names, website structure, or content delivery mechanisms may gradually reduce the effectiveness of static classifiers. As IMSS operators continually adapt to avoid being blocked or identified, our classifier might require periodic updates to maintain high detection accuracy. Additionally, newly emerging IMSS sites may present different characteristics that our current method does not account for, highlighting the need for ongoing monitoring and refinement of detection strategies. Secondly, our tracking blocklist-based detection approach may be incomplete. We rely on *EasyList*, *EasyPrivacy*, *adservers*, and *Disconnect* blocklists to detect tracking cookies and fingerprinting techniques. These lists are popular tracking lists that are well-known and used by both end-users and as ground-truth in academic works [12, 16, 20, 57]. Although we believe this limitation does not significantly affect our measurements and that our work still provides a comprehensive view of various aspects of online tracking on IMSS, we plan to address it in future iterations. Specifically, we intend to integrate a machine learning-based tracker blocker and leverage code structure abstraction using Abstract Syntax Trees to improve detection across diverse linguistic and regional patterns. Next, we limited our implementation to English websites only. Upon trying to expand to other languages, we found that the HTML patterns - that our IMSS identifier is dependent on - differ between different languages, which we intend to improve in future work. Finally, our study primarily focuses on publisher-specific IDs related to Google's services, a major player in the advertising and analytics ecosystem as in Ref. [46]. While this approach provides substantial coverage of real-world scenarios, it also limits our scope to the Google ecosystem, excluding other significant ad networks and analytics services. By concentrating on Google-specific IDs, we may not fully capture the diversity of tracking practices across different platforms. Our approach might miss common ownership if an administrator assigns different IDs to separate containers. This limitation highlights the potential for overlooked connections between sites that share the same owner but use distinct IDs. Further research is necessary to explore and analyze other ad networks and analytics services, providing a more comprehensive understanding of the broader ecosystem.

6.3 Ethical consideration

This study focuses on identifying and analyzing IMSS for research purposes. It is conducted with the intent to understand the ecosystem and the online tracking mechanisms used by these sites to evade detection and to contribute to broader efforts in combating digital piracy. All data collection and analysis are performed with the utmost respect for legal and ethical standards. Additionally, we are mindful of the potential risks associated with engaging with illegal content. Our interactions with IMSS sites are strictly for research purposes, and we have avoided any actions that could be perceived as endorsing or enabling the distribution of pirated

[11] https://almanac.httparchive.org/en/2021/privacy#fingerprinting

[12] https://developers.google.com/privacy-sandbox/private-advertising/topics

material. The research aims to support anti-piracy measures while avoiding any promotion or facilitation of illegal activities.

7 CONCLUSION

In this paper, we explored the online tracking schemes on IMSS. We implemented a method to detect IMSS sites, achieved a recall of 84.31%, and identified 283 new IMSS-hosting websites on Tranco Top 1M domains. Our analysis showed that a select few IMSS sites attract considerable attention, often employing complex redirection patterns to be able to evade blocklisting by simply changing the redirection chain. We also uncovered 11 cases of co-ownership by analyzing Google Identifiers, where multiple sites share the same identifiers, indicating common operation. Our investigation across seven vantage points revealed that over 95% of IMSS sites include at least one third-party tracker, with lower tracker presence observed in EU countries. Moreover, we found that fingerprinting, rather than third-party cookies, is the predominant tracking method used on these sites. These findings underscore the pervasive nature of tracking on IMSS sites and the sophisticated techniques employed to evade detection and maintain operation. The widespread use of fingerprinting highlights a significant challenge, as this technique is more invasive and difficult to mitigate. To foster reproducibility and further research, we have made our tools, datasets, and analysis scripts publicly available at [54].

ACKNOWLEDGMENTS

This research was supported by the Max Planck Society. We thank Nicolas Steinacker-Olsztyn and Huy Pham for their assistance in correcting typographical errors and grammatical issues. We also appreciate the anonymous PoPETs Editor and Reviewers for their valuable feedback, which significantly improved the paper.

REFERENCES

[1] 2021. *Breaking (B)ads: How Advertiser-Supported Piracy Helps Fuel A Booming Multi-Billion Dollar Illegal Market.* Technical Report. https://www.digitalcitizensalliance.org/clientuploads/directory/Reports/Breaking-Bads-Report.pdf Accessed August 31, 2024.

[2] 2023. alliance4creativity.com. https://www.alliance4creativity.com/wp-content/uploads/2023/12/WDWK-2022-worldwide-071223.pdf. Accessed August 31, 2024.

[3] @4saferinternet. 2015. digitalcitizensalliance.org. https://www.digitalcitizensalliance.org/clientuploads/directory/Reports/goodstillbad.pdf. Accessed August 31, 2024.

[4] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the ACM CCS*. 674–689.

[5] LGPD Et Al. 2018. General Personal Data Protection Act (LGPD). https://lgpd-brazil.info/. Accessed Agust 31, 2024.

[6] France Belanger, Janine S Hiller, and Wanda J Smith. 2002. Trustworthiness in electronic commerce: the role of privacy, security, and site attributes. *The journal of strategic Information Systems* 11, 3-4 (2002), 245–270.

[7] Frédéric Besson, Nataliia Bielova, and Thomas Jensen. 2013. Hybrid information flow monitoring against web tracking. In *Proceedings of the IEEE CSF 2013*. IEEE, 240–254.

[8] David Blackburn, Jeffrey Eisenach, and David Harrison. 2019. *Impacts of Digital Video Piracy on the U.S. Economy.* Report. Motion Picture Association.

[9] Károly Boda, Ádám Máté Földes, Gábor György Gulyás, and Sándor Imre. 2012. User tracking on the web via cross-browser fingerprinting. In *Information Security Technology for Applications*. Springer, 31–46.

[10] Ed Chau and Robert Hertzberg. 2018. California Consumer Privacy Act. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375. Accessed August 31, 2024.

[11] Peggy E Chaudhry. 2017. The looming shadow of illicit trade on the internet. *Business Horizons* 60, 1 (2017), 77–89.

[12] Quan Chen, Peter Snyder, Ben Livshits, and Alexandros Kapravelos. 2020. Improving Web Content Blocking With Event-Loop-Turn Granularity JavaScript Signatures. *arXiv preprint arXiv:2005.11910* (2020).

[13] European Commission. 2018. The General Data Protection Regulation (GDPR). https://gdpr-info.eu/. Accessed August 31, 2024.

[14] Dana Dahlstrom, Nathan Farrington, Daniel Gobera, Ryan Roemer, and Nabil Schear. 2006. Piracy in the digital age. *History of Computing* (2006), 1–24.

[15] Savino Dambra, Iskander Sanchez-Rola, Leyla Bilge, and Davide Balzarotti. 2022. When Sally met trackers: Web tracking from the users' perspective. In *Proceedings of the USENIX Security*. 2189–2206.

[16] Ha Dao. 2022. *Detection, characterization, and countermeasure of first-party cooperation-based third-party web tracking.* Ph. D. Dissertation. The Graduate University for Advanced Studies.

[17] Disconnect. 2019. *Tracker Descriptions for Fingerprinters and Cryptominers.* Retrieved August 22, 2024 from https://github.com/disconnectme/disconnect-tracking-protection/blob/master/descriptions.md Accessed August 31, 2024.

[18] Ana Durrani. 2024. Top Streaming Statistics In 2024 — forbes.com. https://www.forbes.com/home-improvement/internet/streaming-stats/. Accessed August 31, 2024.

[19] Peter Eckersley. 2010. How unique is your web browser?. In *Proceedings on PoPETs*. Springer, 1–18.

[20] Steven Englehardt, Jeffrey Han, and Arvind Narayanan. 2018. I never signed up for this! Privacy implications of email tracking. (2018), 109–126.

[21] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the ACM CCS*. 1388–1401.

[22] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the WWW*. 289–299.

[23] fanboy, MonztA, Famlam, and Khrin. 2005. *EasyList.* Retrieved August 22, 2024 from https://easylist.to/easylist/easylist.txt Accessed August 31, 2024.

[24] fanboy, MonztA, Famlam, and Khrin. 2005. *EasyPrivacy.* Retrieved August 22, 2024 from https://easylist.to/easylist/easyprivacy.txt Accessed August 31, 2024.

[25] Imane Fouad, Nataliia Bielova, Arnaud Legout, and Natasa Sarafijanovic-Djukic. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. In *Proceedings on PoPETs*.

[26] Jonathan Frost. 2017. Global piracy increases throughout 2017, MUSO reveals — muso.com. https://www.muso.com/magazine/global-piracy-increases-throughout-2017-muso-reveals. Accessed August 31, 2024.

[27] Matthias Gotze, Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. 2022. Measuring web cookies in governmental websites. In *Proceedings of the ACM WebSci*. 44–54.

[28] Luke Hsiao and Hudson Ayers. 2019. The Price of Free Illegal Live Streaming Services. arXiv:1901.00579 [cs.CR] https://arxiv.org/abs/1901.00579

[29] Damilola Ibosiola, Benjamin Steer, Alvaro Garcia-Recuero, Gianluca Stringhini, Steve Uhlig, and Gareth Tyson. 2018. Movie pirates of the caribbean: Exploring illegal streaming cyberlockers. In *Proceedings of the AAAI ICWSM*, Vol. 12.

[30] Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat* 37 (1901), 547–579.

[31] Justdomains. 2022. DOMAIN-ONLY Filter Lists. https://github.com/justdomains/blocklists. Accessed August 31, 2024.

[32] Sina Keshvadi and Carey Williamson. 2021. An empirical measurement study of free live streaming services. In *Proceedings of the PAM*. Springer, 111–127.

[33] Maciej Korczynski, Samaneh Tajalizadehkhoob, Arman Noroozian, Maarten Wullink, Cristian Hesselman, and Michel Van Eeten. 2017. Reputation metrics design to improve intermediary incentives for security of TLDs. In *Proceedings of the IEEE EuroS&P*. IEEE, 579–594.

[34] Balachander Krishnamurthy and Craig Wills. 2009. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of WWW*. 541–550.

[35] IAB Tech Lab. 2019. IAB Tech Lab ads.txt Specification Version 1.0.2. https://iabtechlab.com/wp-content/uploads/2019/03/IAB-OpenRTB-Ads.txt-Public-Spec-1.0.2.pdf. Accessed August 31, 2024.

[36] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, and Wouter Joosen. 2019. TRANCO: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the NDSS*.

[37] Chung Hun Lee and David A Cranage. 2011. Personalisation–privacy paradox: The effects of personalisation and privacy assurance on customer responses to travel Web sites. *Tourism Management* 32, 5 (2011), 987–994.

[38] Timothy Libert. 2015. Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on One Million Websites. *International Journal of Communication* (2015).

[39] Peter Lowe. 2001. *Blocklist for use with hosts files to block ads, trackers, and other nasty things.* Retrieved August 22, 2024 from https://pgl.yoyo.org/adservers/serverlist.php?hostformat=hosts&showintro=1&mimetype=plaintext Accessed August 31, 2024.

[40] Jonathan R Mayer and John C Mitchell. 2012. Third-party web tracking: Policy and technology. In *Proceedings of the IEEE S&P*. IEEE, 413–427.

[41] Johan Mazel, Richard Garnier, and Kensuke Fukuda. 2019. A comparison of web privacy protection techniques. *Computer Communications* 144 (2019), 162–174.

[42] Mozilla. 2005. Public Suffix List. https://publicsuffix.org/. Accessed August 31, 2024.

[43] Alexios Nikas, Efthimios Alepis, and Constantinos Patsakis. 2018. I know what you streamed last night: On the security and privacy of streaming. *Digital Investigation* 25 (2018), 78–89.

[44] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. 2015. Privaricator: Deceiving fingerprinters with little white lies. In *Proceedings of WWW.* 820–830.

[45] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2013. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of the IEEE S&P.* IEEE, 541–555.

[46] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2022. Leveraging google's publisher-specific ids to detect website administration. In *Proceedings of the ACM Web Conference.* 2522–2531.

[47] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2023. Who funds misinformation? A systematic analysis of the ad-related profit routines of fake news sites. In *Proceedings of the WWW.* 2765–2776.

[48] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 2019. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *Proceedings of WWW.* 1432–1442.

[49] M Zubair Rafique, Tom Van Goethem, Wouter Joosen, Christophe Huygens, and Nick Nikiforakis. 2016. It's Free for a Reason: Exploring the Ecosystem of Free Live Streaming Services. In *Proceedings of the NDSS.*

[50] Ali Rasaii, Shivani Singh, Devashish Gosain, and Oliver Gasser. 2023. Exploring the cookieverse: A multi-perspective analysis of web cookies. In *Proceedings of the PAM.* Springer, 623–651.

[51] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and defending against third-party tracking on the web. In *Proceedings of USENIX NSDI.* 155–168.

[52] Sebastian Schelter and Jérôme Kunegis. 2016. Tracking the trackers: A large-scale analysis of embedded web trackers. In *Proceedings of the International AAAI Conference on Web and Social Media.* 679–682.

[53] Scrapinghub. 2013. *Python parser for Adblock Plus filters.* Retrieved August 22, 2024 from https://github.com/scrapinghub/adblockparser Accessed August 31, 2024.

[54] Hussein Sheaib, Anja Feldmann, and Ha Dao. 2024. Tool, raw dataset, and analysis scripts for Illegal Movie Streaming Services measurement. https://doi.org/10.17617/3.STVMDI

[55] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke. 1996. State of the information privacy literature: Where are we now and where should we go? *MIS Quarterly* 20, 2 (1996), 167–196.

[56] Michael D Smith and Rahul Telang. 2009. Competing with free: The impact of movie broadcasts on DVD sales and Internet piracy. *Mis Quarterly* (2009), 321–338.

[57] Oleksii Starov and Nick Nikiforakis. 2018. Privacymeter: Designing and developing a privacy-preserving browser extension. In *Proceedings of the ESSoS.* Springer, 77–95.

[58] Rahul Telang. 2018. Does online piracy make computers insecure? evidence from panel data. *Evidence from Panel Data (March 12, 2018)* (2018).

[59] Karyn A. Temple. 2023. Testimony before the U.S. House of Representatives Committee on the Judiciary, Subcommittee on Courts, Intellectual Property, and the Internet: Hearing on "Digital Copyright Piracy: Protecting American Consumers, Workers, and Creators". Testimony of Senior Executive Vice President and Global General Counsel of Motion Picture Association, Inc.. Accessed August 31, 2024.

[60] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Measuring the impact of the gdpr on data sharing in ad networks. In *Proceedings of ACM CCS.* 222–235.

[61] Hao Yang, Kun Du, Yubao Zhang, Shuang Hao, Zhou Li, Mingxuan Liu, Haining Wang, Haixin Duan, Yazhou Shi, Xiaodong Su, et al. 2019. Casino royale: a deep exploration of illegal online gambling. In *Proceedings of the ACSAC.* 500–513.

A IMSS SITE PATTERNS

The detailed explanation for the eight additional patterns to build our classifier is presented in Figure 13. Each pattern is labeled from (a) to (h), and they represent different hierarchical structures of HTML elements commonly found on the IMSS websites.

B SLD WITH MULTIPLE TLDs

Table 5 shows the second-level domains with more than one TLD. The domains are grouped by the number of TLDs associated with each second-level domain. It provides insights into the reuse or distribution of second-level domains across different TLDs.

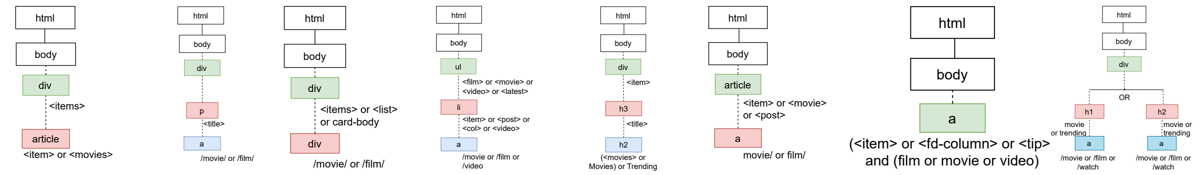
# TLDs	Second-Level Domains
2	fmovie, 6movies, attackertv, rivestream, onionplay, 0123movie, 123moviestv, animeworld, 123movieshub, 1hd, himovies, movie4u, myflixerz, flixwave, watchseries, 123series, cinego, doramasflix, binged, new-123movies, flixhd, 1movies, cataz, soap2dayhd, 123serieshd, f2movies (**26**)
3	123-movies, 123movies123, ymovies, putlockers, gomovies, hurawatch, flixhq, sflix, cineb, solarmovies, primewire, moviesjoy, theflixer, hdtoday (**14**)
4	soap2day, dopebox, yesmovies, movies2watch, pelisflix (**5**)
5	bflix, movies123 (**2**)
6	solarmovie (**1**)
7	myflixer, 123moviesfree, pelisflix2, 123movies (**4**)
8	fmovies (**1**)
9	putlocker (**1**)
10	lookmovie (**1**)

Table 5: Second-level domains with more than one TLD, grouped by the number of TLDs.

C UNIQUE TRACKERS BY LOCATIONS

The following is a list of tracker domains that are observed exclusively in specific locations. These domains are unique to their respective regions, meaning they are not detected in other geographical VPs. This specificity can indicate regional targeting or localization strategies by these trackers, which may be tailored to the preferences or behaviors of users in these particular VPs.

- **Australia**: adtdp.com, metatrckpixel.com, admeme.net, idnyfbtt.top, amgdgt.com, presage.io, google.com.au, zucks.net, gsspat.jp, bsdhcoiutdsd.top, ladsp.com, zeekaihu.net (12).
- **Brazil**: bfmio.com, adstanding.com, google.com.br, technoratimedia.com, audrte.com, geoedge.be, rtactivate.com, resetdigital.co (eight).
- **Germany**: eegheecog.net, google.de, nextmillmedia.com (3).
- **India**: b2c.com, lator308aoe.com, rqohrdsbt.top, chirtakautoa.xyz, google.co.in, adpartner.pro, pbstck.com (seven).
- **South Africa**: eacdn.com, labourattention.com, realuniverse24.com, onclickperformance.com, crazyegg.com, facebook.com, taptapnetworks.com, oracleinfinity.io, de17a.com, serving-sys.com, everestads.net, zaltaumi.net, google.co.za (13).
- **Sweden**: creative-bars1.com, yourwebbars.com, google.se, phoawhoax.com (four).
- **The United States**: swoop.com, mdhv.io, petchesa.net, rtbiq.com, videoamp.com, media6degrees.com, admission.net, commander1.com, ispot.tv, zdbb.net, tidaltv.com, storygize.net, swpsvc.com, innovid.com, prfct.co, oatchoagnoud.com, akamaihd.net, survata.com, apolloprogram.io, adsymptotic.com, mmtro.com, rkdms.com, d41.co, globalwebindex.net, clrstm.com, connatix.com, samplicio.us, fzlnk.com, extend.tv, acxiomapac.com, tvpixel.com, jivox.com, sundaysky.com, securedvisit.com, geniusmonkey.com, clinch.co, viatepigan.com, ccgateway.net (38).


html
body
div
<items>
article
<item> or <movies>
html
body
div
p
<title>
a
/movie/ or /film/
html
body
div
<items> or <list> or card-body
div
/movie/ or /film/
html
body
ul
<film> or <movie> or <video> or <latest>
li
<item> or <post> or <col> or <video>
a
/movie or /film or /video
html
body
div
<item>
h3
<title>
h2
(<movies> or Movies) or Trending
html
body
article
<item> or <movie> or <post>
a
movie/ or film/
html
body
a
(<item> or <fd-column> or <tip> and (film or movie or video)
html
body
div
OR
h1
h2
movie or trending
movie or trending
a
a
/movie or /film or /watch
/movie or /film or /watch

(a) Pattern 2 **(b) Pattern 3** **(c) Pattern 4** **(d) Pattern 5** **(e) Pattern 6** **(f) Pattern 7** **(g) Pattern 8** **(h) Pattern 9**

Figure 13: Eight additional patterns used to build IMSS classifier.

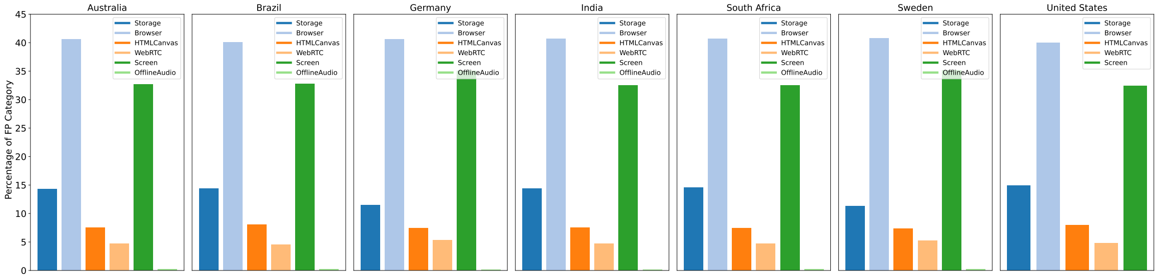


Figure 14: Fingerprinting techniques in IMSS sites across different VPs.

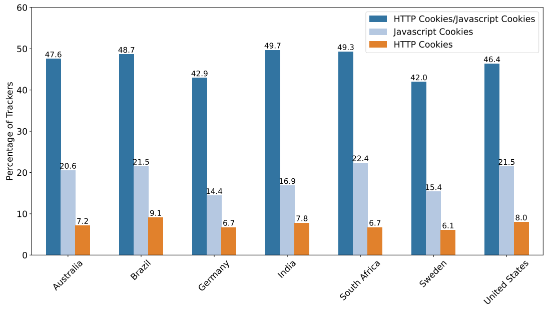


Figure 15: Percentage of trackers that set cookies via HTTP Cookies and JavaScript across different VPs.

D JAVASCRIPT COOKIES AND HTTP COOKIES

Figure 15 shows the percentage of trackers that set JavaScript cookies and HTTP cookies across all locations and over the 10 runs. To determine whether these trackers primarily rely on cookies as their tracking technique, we analyzed all tracker domains from our database. The results show that, on average, more than 60% of trackers set cookies in some form at every location. Specifically, over 41% of trackers set both JavaScript and HTTP cookies. Additionally, at least 14.4% of trackers set only JavaScript cookies, while at least 6.7% of trackers set only HTTP cookies.

E FINGERPRINTING TECHNIQUES IN IMSS SITES

Figure 14 shows that *BrowserInfo* is the most popular fingerprinting type used by trackers on IMSS sites. Moreover, we find that *ScreenInfo* is the second most used fingerprinting type in all locations and that *OfflineAudioInfo* is the least used in all locations as well.

Fingerprinting severity levels	# Trackers	# Site presence
FingerprintingGeneral (FG)	14	281
FingerprintingInvasive (FI)	10	65

Table 6: Presence of fingerprinting severity levels.

In addition, *Disconnect* also categorizes fingerprinting activities into two severity levels: *FingerprintingInvasive* (FI) and *FingerprintingGeneral* (FG). For FI, *Disconnect* conducts a detailed assessment of each service's fingerprinting activities using various techniques. This includes analyzing results from API monitoring studies to detect access to sensitive fingerprinting attributes, such as canvas elements or audio context. They also identify the use of specific fingerprinting libraries, like *fingerprints.js*[13], to classify a service as FI. In contrast, services labeled as FG are categorized based on a review of their privacy policies. This categorization focuses on whether a service's privacy policy indicates the potential for general fingerprinting activities, without the more detailed technical analysis that characterizes the FI category. To further illustrate this categorization, Table 6 presents a detailed breakdown of these fingerprinting categories, showing both the number of trackers classified as FG and FI, as well as the number of IMSS sites on which these FG and FI trackers are present. In addition to the 14 distinct FG-classified trackers found on 281 IMSS sites, we even identified 10 distinct FI trackers present on 65 sites. The presence of both FI and FG trackers on IMSS sites highlights the range of fingerprinting practices employed within this ecosystem, from invasive to more generalized methods, underscoring the varying levels of privacy risk for users accessing these sites.

[13] https://github.com/fingerprintjs/fingerprintjs