Tales from the Porn: A Comprehensive Privacy Analysis of the Web Porn Ecosystem

Pelayo Vallina
IMDEA Networks Institute /
Universidad Carlos III de Madrid
pelayo.vallina@imdea.org

Álvaro Feal
IMDEA Networks Institute /
Universidad Carlos III de Madrid
alvaro.feal@imdea.org

Julien Gamba
IMDEA Networks Institute /
Universidad Carlos III de Madrid
julien.gamba@imdea.org

Narseo Vallina-Rodriguez
IMDEA Networks Institute / ICSI
narseo.vallina@imdea.org

Antonio Fernández Anta
IMDEA Networks Institute
antonio.fernandez@imdea.org

ABSTRACT

Pornographic websites are among the most visited web pages globally. Because of the sensitive content that they offer, modern privacy regulations try to control user tracking activities in such sites. However, little is known about the privacy risks that users face when visiting such sites and their regulatory compliance. In this paper, we present the first comprehensive and large-scale analysis of 6,843 porn websites. We provide an exhaustive behavioral analysis of the use of tracking methods by these services, their concerning lack of encryption and cookie consent forms, and their lack of regulatory compliance, including the presence of age-verification mechanisms. The results indicate that, as in the regular web, tracking is prevalent across pornographic sites: 72% of the websites use third-party cookies and 5% leverage advanced fingerprinting scripts. Further, our analysis reveals a third-party tracking ecosystem semi-decoupled from the regular web in which various analytics and advertising services track users across, and outside, the porn web. We conclude our paper with a regulatory compliance analysis in the context of the EU General Data Protection Regulation (GDPR), and newer legal requirements to implement access control mechanisms (*e.g.,* UK's Digital Economy Act). We find that only 16% of the analyzed websites have a privacy policy and only 4% have a cookie consent banner. The use of verifiable access control mechanisms is limited to prominent pornographic websites.

1 INTRODUCTION

Pornographic (porn) websites are among the most visited and also lucrative online services since the early days of the World Wide Web [86]. Pornhub, the most visited porn website according to Alexa's domain rank [4], had 33.5 Billion visits and was returned in 30.3 Billion web searches in 2018 [67]. MindGeek, Pornhubs' parent company, has reported over half a billion dollars of revenue during the 2015 fiscal year [52].

Modern privacy regulations like the EU General Data Protection Regulation (GDPR) [29] and California's Consumer Privacy Act (CCPA) [16] consider sexual information of an individual as highly sensitive data. Most privacy regulations also require organizations with an online presence to request informed consent from users prior to any data collection [16, 29, 40, 42]. However, as in the case of regular websites, pornographic ones also integrate third-party advertising and analytics services with the capacity to track users' interaction with such services and, therefore, potentially infer a visitor's sexual orientation. The collection of this information, in addition to the absence of secure network protocols like HTTPS, could put at risk visitors of those websites, specially those connecting from countries where certain sexual orientations are prosecuted [14, 35, 38, 72, 79].

Despite the many research efforts that took place in the last decade to identify and quantify the presence and use of tracking technologies in the web, no study has deep dived yet into the privacy risks of sensitive websites, like pornographic ones. It is unclear, as a result, whether pornographic websites can pose a privacy risk to their visitors and if they comply with existing privacy regulations and newer rules to control minor's access to adult content, like the UK's Digital Economy Act from 2017 [49]. Anecdotal evidence suggests that there are significant differences between the third-party organizations operating in the porn and the regular web tracking industry [9]. In fact, large online ad networks such as Google Ads set strict constraints for porn-related publishers, prohibiting the advertising of adult products and services [35]. These restricting terms of services – possibly driven by fear of damaging their brand reputation – opened new market opportunities for other actors who have specialized in providing advertising and tracking technologies to adult sites. This context has created, as a result, a parallel ecosystem of third-party service providers in the porn

ecosystem who has not been scrutinized by regulators, policy makers, and the research community.

In this paper, we develop and use a methodology to perform the first holistic analysis of pornographic websites from a privacy and transparency perspective. Our main contributions are:

- We design a semi-supervised method to compile a representative sample of pornographic websites using publicly available resources (Section 3). After manually inspecting and removing false-positives, we collect 6,843 porn websites.
- We develop and use a methodology to study the presence of third-party services in the porn ecosystem (Section 4). We compare third/-party service's behavior and presence in pornographic websites with those present in the most popular web sites according to Alexa's rank. We find 3,673 third-party services embedded in porn websites, including companies specialized in porn websites (*e.g.,* ExoClick) and well-known advertising companies (*e.g.,* DoubleClick), analytics services (*e.g.,* GoogleAnalytics), and also domains associated to data brokers (*e.g.,* Acxiom). 85% of the third-party services embedded on porn websites, do not appear on the regular web ecosystem.
- We study the behavior of pornographic websites and the third-party tracking services embedded in them (Section 5). We find the presence of third-party HTTP Cookies in 72% of the analyzed pornographic websites, while 5% of them also use advance fingerprinting techniques to identify visitors uniquely. 91% of the scripts we found using canvas fingerprinting are not indexed by EasyList and EasyPrivacy.
- We quantify behavioral differences on porn websites depending on the user's location and jurisdictional area (Section 6). We conclude that the number of third-party services is quite stable across countries, yet there are regional third-party services that only operate in specific regions: *e.g.,* 27 ATS services only appear in Russia.
- We develop and validate a method to automatically analyze the transparency and regulatory compliance of pornographic websites (Section 7). Specifically, we study the presence of cookie consent banners, privacy policies, and age-verification mechanisms. Our analysis reveals a significant absence of privacy policies and consent forms across pornographic websites in spite of their sensitivity. This pattern holds even in regions with strict regulatory frameworks: only 16% of the websites have privacy policies when accessed from the EU. Further, the use of cookie consent banners on porn websites is lower than in the regular web, since only 4% of the analyzed porn websites implement cookie consent forms.

Our study reveals a concerning lack of transparency in pornographic websites despite a significant presence of third-party trackers embedded in them and an increasing regulatory pressure. Therefore, we believe that our study will contribute to stress the importance of studying sensitive subsets of the world wide web in depth. This type of effort is not only needed to effectively inform the privacy debate but also to promote user awareness.

2 BACKGROUND

The many privacy abuses inflicted by the online industry in the latest decades have motivated regulatory and legislative efforts to protect consumers' privacy and digital rights in the web. New comprehensive data protection laws such as the European General Data Protection Regulation (GDPR) – which became effective on May 25th, 2018 [18]– aim to bring transparency to web services and empower users with control over their personal identity in the web and beyond. In the case of online services, this objective is achieved by forcing companies to obtain explicit consent from any European visitor before collecting, processing, or sharing personal data. The GDPR, which will be complemented by the ePrivacy directive in 2019 [17, 27], also gives users the right to access, correct, and delete their data collected and stored by online services, revoke their collection consent at any time, and object to automatic data processing.

In the context of pornographic websites, the GDPR imposes additional requirements and restrictions on data collectors and controllers due to the sensitive nature of pornographic services. In fact, article 9.1 [29] of the GDPR states that *"processing of data concerning a natural person's sex life or sexual orientation shall be prohibited."* Until the ePrivacy regulation becomes effective, the GDPR will require website owners to obtain explicit consent from users to install and use tracking methods, such as HTTP cookies, but when it will be strictly required to provide a service requested by the user, legal mandate, or to carry out certain transmissions [28].

Similar regulatory efforts are taking place in other jurisdictions that used to have a traditional *laissez faire* attitude towards privacy. Notable examples are California's Consumer Privacy Act [16] (CCPA), (passed in June 2018), the Japanese Act on Protection of Personal Information [42] (effective since May 2017) and the Indian Personal Data Protection Bill of 2018 (PDP) [40]. All the aforementioned regulations classify and consider information regarding a user's sexual life and orientation as sensitive personal data that require special treatment.

2.1 Access Control in Pornographic Sites

Several countries have passed laws to enforce the deployment of age verification mechanisms to block minors from

accessing pornographic material. The 2017 Digital Economy Act [49] in the United Kingdom is one remarkable example that will finally become effective on July 15th, 2019 [87]. To comply with these new age verification laws, the industry had to design and develop tools such as AgeID, a verifiable age-verification tool developed by MindGeek [8], the company that owns some of the most prominent pornographic websites. This tool is expected to become an standard that will be used across all porn sites in the UK [80]. Other regions in the world have followed more drastic strategies. The Russian government required Pornhub users to login with a social network profile —linked to their passport number [14, 88]. This measure has raised several ethical and privacy concerns. A complementary effor to the aforementioned methods is the proposal made by the Association of Sites Advocating Child Protection (ASACP) [11]. This non-for-profit organization has created a Restricted-for-Adults (RTA) meta tag to assist parents to prevent their children from accessing pornographic material. The fact that there are companies from the online porn industry among the members [10] of this association is considered as a good example of collaboration between the porn industry and external organizations to increase safety and regulatory compliance. Finally, several world countries (*e.g.,* most Middle East countries, India, Iran, or China) actively ban, prosecute, and prohibit access to pornographic content altogether [63, 84]. The 2013 Anti Pornography Act in Uganda prosecutes the broadcasting and trading of pornography [41], while the Anti-Homosexuality Bill Act in 2014 prosecutes LGBTI communities [72].

3 DATA COLLECTION AND METHOD

The first challenge in our study is compiling a representative list of pornographic websites, beyond the main providers. For that, we implement a semi-supervised approach that combines three different data sources and steps with varying levels of accuracy:

(1) We aggregate all the pornographic websites indexed by three websites specialized in aggregating, recommending, and classifying pornographic content [2, 3, 59]. This process provided us with 342 porn websites.
(2) We extract 22 websites classified as Adult sites by Alexa's website categorization service [82].
(3) We identify 7,735 websites potentially offering pornographic content by searching for keywords related to pornographic and adult content in URLs indexed by the web-oriented Alexa domain rank [4] such as "porn","tube", "sex","gay","lesbian" and "mature". 90% of the porn websites in our corpus contains at least one of these keywords on their URL.

The combination of these three methods allowed us to identify 8,099 `potential` pornographic websites. However, the third string-based method could introduce false positives, since we manually chose a set of words that are highly but not only related to pornographic material (*e.g.,* PornTube offers pornographic content while YouTube is not). To identify and remove false positives, we implement a purpose-built crawler to download their content (DOM and screenshots) which was then manually inspected. In total, we found 1,256 false positives, many of which were unpopular websites that were not responsive at the time of running the crawler. We investigate below the stability of these domains. After this sanitation process, we obtained a corpus of 6,843 pornographic websites of various kinds, including websites hosting user-uploaded videos and live streaming content, or websites acting as proxies to pornographic material (*e.g.,* pornsource.com) among many others. Finally, we also built a reference dataset containing 9,688 popular non-pornographic websites[1] to study the commonalities and differences between sensitive pornographic websites and regular ones.

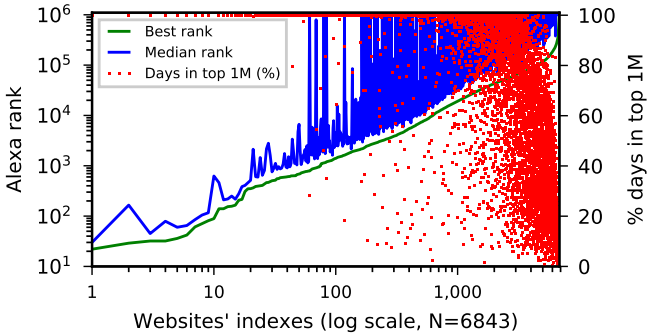

Figure 1: Best (green) and median (blue) Alexa rank and percentage of days a pornographic website is present in the top-1M throughout 2018. Websites are ordered in the x-axis by their highest rank.

Popularity of Pornographic Websites: We use the Alexa top-1M as a proxy to measure the stability, popularity, and representativeness of our corpus of pornographic websites. Figure 1 shows the highest and median rank value for each one of the identified porn websites, as well as the percentage of days each website was in the Alexa top-1M over the year 2018. [2] We found that 1,103 websites (16%) were always present in the Alexa top-1M, and just 16 of them were always within the top-1K websites for the whole year. This includes famous pornographic websites such as `pornhub.com`, `xvideos.com` or `livejasmin.com`.

[1] We extracted the websites from Alexa's top-10K (10^{th} of January 2019).
[2] We consider their popularity for a whole year in order to account for any eventual bias caused by the Alexa ranking [76].

3.1 Web Crawlers

Our analysis and data collection workflow uses two different crawlers to empirically study the behavior of pornographic websites as shown in Figure 2. We only crawl the landing page of websites (or the page right after the age verification mechanism if applicable) to avoid generating advertising revenues and accessing specific content. Therefore, our study presents a lower-bound estimation of the presence of trackers in pornographic websites.

OpenWPM: Rather than implementing our own crawler, we use OpenWPM [24] because of its simplicity, stability, and the versatility of the features that it offers. OpenWPM is based on Firefox version 52 and allows collecting all the HTTP and HTTPS requests and responses generated while crawling a website, including those created by embedded third-party services such as advertising and tracking services. We extend it to also extract the chain of requests caused by Real-Time Bidding (RTB) processes (*i.e.,* the inclusion chain) [12] to control and reduce any bias introduced by RTB in the number of third-party services embedded in the target websites [12]. For that, we analyze the HTTP `Referrer` headers and remove the third-parties not directly called by the publisher. Furthermore, OpenWPM provides support for detecting different tracking methods and monitoring JavaScript API calls triggered by each website, many of which are used for fingerprinting users [24]. We use the same browser session —we don't close and open the browser between visits — for the duration of the crawling process, in order to be able to capture cookie synchronizations (Section 5.1.2). [3] It is also important to note that we only crawl each page once, giving us a lower bound on tracking activities [31]. We extended OpenWPM to automatically record both HTTP cookie policies and consent forms, so that we can study the regulatory compliance for each pornographic website in Section 7.

Selenium: We implement a second purpose-built Selenium-based Chrome crawler to (1) detect and bypass age-verification mechanisms in pornographic sites; and (2) fetch their privacy policies when available. We separate this data collection process from OpenWPM crawls to avoid any instrumentation bias introduced by the need of interacting with each website in order to identify their privacy policies. To detect and quantify the support for age verification mechanisms, our crawler parses the landing page of a website and searches for floating elements and the words "Yes", "Enter", "Agree", "Continue" and "Accept" in 8 languages[4]. We choose these keywords after a manual inspection of part of the websites in our corpus. Furthermore, to remove false-positives introduced by using string base matching, our crawler also looks for the parent and grandparent elements in the HTML DOM to verify the presence of warning messages to users about the type of content embedded in the webpage. If this message is found, then the crawler clicks on the element in order to access the landing page. We then fetch privacy policies by searching for URL links containing the keywords "Privacy" and "Policy" in the same 8 languages. We manually validate the accuracy of our method in Section 7.2.

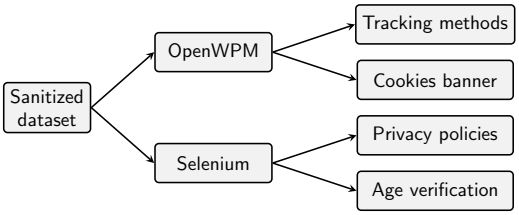

Figure 2: Workflow of our data collection techniques.

Geographical diversity: One of the goals of this work is to study whether pornographic websites behave differently depending on the user location and jurisdiction. To answer that, we run our crawls from a vantage point located in Spain, leveraging two commercial VPN providers — NordVPN [61] and PrivateVPN [68][5] — to gain access to vantage points in other EU members, Singapore, India, Russia, USA, and the UK. [6] When crawling from Russia and India, we could not access 21 and 168 pornographic websites, respectively. We can not assert whether this is due to country-level censorship or server-side blocking [55].

4 THE PORN WEB ECOSYSTEM

As of today, the research community lacks of generalizable and robust methods to classify domains by their purpose and role, and to identify the parent company [71]. However, gaining this knowledge is critical, not only to identify the organization providing the service, but also for accountability aspects like when they are tracking users in sensitive websites. In this section we explore (1) the main stakeholders providing pornographic services in the web (Section 4.1); and (2) the analysis of third-party advertising and tracking services embedded in porn sites (Section 4.2).

4.1 Discovering Website Owners

We start this analysis by crawling the landing page and privacy policy (when available) of each pornographic website,

[3] We establish a timeout of 120s for loading a website in order to prevent our crawlers from becoming stagnant.

[4] We select English, Spanish, French, Portuguese, Russian, Italian, German, and Romanian for being the most common languages in pornographic websites.

[5] We select those VPN providers because 1) they do not appear to manipulate traffic according to our experiments, and 2) they forward traffic through VPN servers rather than through real users in a P2P fashion [46].

[6] We perform these measurements in the U.K. to study websites' compliance with the Digital Economy Act [49])

to search for organization-level details. However, for the majority of pornographic websites, this information is either vague or incomplete: *e.g.,* some websites only report a postal address rather than a company name accompanied by legal information. To increase the coverage of our crawler, we augment the data directly fetched from pornographic websites with DNS, WHOIS, and X.509 certificate information, as well as insights obtained from white papers, scientific articles, and public reports about the pornographic industry [34, 81].

The combination of these methods only allowed us to identify 16 companies owning 191 pornographic websites. We could not identify any reliable corporate information for 97% of the websites in our corpus. This lack of corporate or organizational transparency is particularly concerning for these sites engaging in user tracking and embedding third-party services —further discussed in Sections 5 and 7— as their visitors will not be able to effectively exercise their privacy rights (*e.g.,* demanding access, corrections, and deletion of their data as indicated in the GDPR).

Main pornographic website operators: Table 1 shows the 10 largest clusters of pornographic websites ordered by the number of individual websites that they own. These companies own and operate 3% of the total websites in our corpus. Typically, these clusters of pornographic websites are created through acquisitions and mergers between companies, similar to the industry trends present in the online advertising and tracking industry [60, 71]. In fact, pornographic websites are typically federated. This gives them the ability to reach out larger audiences and increase advertising revenues through affiliated services, while also re-publishing and sharing pornographic material across sites. As opposed to previous studies, we will analyze any provider of pornographic content, regardless of their popular and market presence. As our results will show, this is critical in order to gain a complete understanding of the ecosystem and its practices.

Monetization Models: The majority of pornographic websites combine different monetization mechanisms, such as online advertising-based models (see Section 4.2) and subscription (premium) services. We perform a semi-automatic classification of these websites to infer their business models. First, we parse the landing page of the websites in our dataset and look for keywords that may indicate the option to create an account (*e.g.,* "Log In", "Sign Up") or "Premium" services. We use this signal as a proxy to identify which websites may offer subscription-based services after authentication. Then, we label the subscription as free or paid by manually inspecting the website. We also verify that the keywords for creating an account and for detecting premium services maintain stable independently of the language of the webpage. Thanks to this method, we can conclude that 14% of the porn websites in our corpus offer subscription options; and only 23% of the websites offering subscriptions are paid.

Table 1: Largest clusters of pornographic sites, grouped by their parent company. For each company, we report the number of individual websites owned and the one with the highest Alexa rank throughout 2018. A larger cluster size does not necessarily translate into popularity.

Company	# sites	Most popular site (rank)
Gamma Entertainment	70	`evilangel.com` (5,301)
MindGeek	39	`pornhub.com` (22)
PMG Entertainment	15	`private.com` (7,758)
Vivid Entertainment	12	`vivid.com` (33,496)
Docler Holding	10	`livejasmin.com` (36)
Liberty Media	7	`corbinfisher.com` (26,436)
Zero Tolerance	6	`ztod.com` (40,676)
JM Productions	6	`jerkoffzone.com` (147,753)
WGCZ	5	`xvideos.com` (32)
Eurocreme	5	`eurocreme.com` (110,012)

Table 2: Number of first-, third-party, and third-party ATS domains found on our dataset of pornographic and regular websites.

Domain category	Pornographic websites (P)	Regular websites (R)	$P \cap R$
Corpus size	6,346	8,511	—
First-party	727	3,852	—
Third-party	5,457	21,128	889
Third-party ATS	663	196	86

While the study of the privacy risks of subscription-based services is outside the scope of this paper, it may be possible that once a user creates an account, all of their actions might be also linked to their profile and banking information.

4.2 Third-Party Services in Porn Websites

A large number of pornographic websites rely on online advertisements to monetize their user base and content. However, large ad networks set strict limitations on the usage of their services in pornographic websites, possibly as a measure to protect their brand reputation [35]. This state of affairs has given birth to lesser known advertisement and tracking services (ATS) specialized in adult content which have escaped research and regulatory scrutiny. We conjecture that our current limited understanding of trackers in sensitive websites has been caused by the low penetration of these trackers across the whole web landscape, hence falling in the long-tail, and also by the fact that many pornographic websites are rarely indexed in domain ranks [76].

In this subsection we study the third-party services and organizations operating in the online porn industry, and compare them with those present in regular websites. With our OpenWPM-based crawler, we identify 3,673 different third-party domains embedded in the set of 6,346 pornographic websites that we could successfully crawl. An eyeball analysis of these domains reveals that the majority of them belong to third-party analytics and advertising services, but also to CDN providers and social networks. To obtain a more accurate picture of the third-party tracking ecosystem in pornographic websites, we use the following complementary heuristics to (1) label and classify domains embedded in pornographic websites as first-, third-party, or third-party advertising and tracking (ATS) services; and (2) attribute hostnames to organizations:

(1) **Third-party service extraction:** We collect all the URLs from each HTTP(S) requests triggered by our OpenWPM-based crawler to identify the presence of third parties. For comparison, we run our crawl for our sets of pornographic and regular websites. For each URL and request, we compare its fully qualified domain name (FQDN) and its X.509 certificate information (when available) along with the FQDN and certificate information of the host website, to determine whether a service is a first or third party. If we cannot establish a relationship between a host website and an embedded service based on the previous method, we compute the similarity between the two FQDNs using the Levenshtein distance [51]. We consider the FQDNs to belong to the same entity if their similarity is higher than 0.7. We manually verified the results and found this method to be accurate. This method also allows us to group together domains such as `doublepimp.com` and `doublepimpssl.com`. Thanks to this technique we can successfully label as third party domains 91% of the 6,017 FQDNs contacted when crawling all the porn and regular websites in our corpus.
(2) **ATS classification:** We rely on EasyList and EasyPrivacy blacklists [22] —downloaded on Jan. 29th, 2019— to identify domains belonging to well-known ATS. However, these blacklists are based on rules that consider the whole URL request (*e.g.,* `bbc.co.uk` is not blacklisted, but `bbc.co.uk/analytics` is). Therefore, we match the full URL provided by OpenWPM with these blacklists to identify actual instances of tracking, but we relax the matching method to the base FQDN domain to identify the presence of third-party ATS organizations [93]. If a URL from a third-party organization on a particular website matches with one of the EasyList or EasyPrivacy rules, we label the domain as an ATS present in that website. Using this method, we can label 12% of the third-party services as ATS.
(3) **Finding the parent company for third-party services:** To better understand the trackers and organizations involved in the ecosystem, it is critical to associate third-party domains to its parent company. We initially considered using Disconnect's domain-to-company mapping [20] but we soon realized that it is quite incomplete. Instead, we leverage organization-level information found in the X.509 certificate of each third-party domain[7], hence improving the coverage of Disconnect's list. For instance, using this method we could attribute to Oracle third-party trackers like `addthis.com` (AddThis) [6] and `bluekai.com` (BlueKai) [15] services. [8] After their process, we could find the parent company for 4,477 (74%) FQDNs, accounting for 1,014 companies, while using Disconnect's list we could only detect 142 of them.

Third parties in regular and porn websites. Table 2 compares the number of third-party domains present in our set of pornographic websites with those present in our reference set of regular websites. This comparative analysis uncovers significant differences. In aggregated terms, we find 21,128 third-party domains (FQDNs) in our set of regular websites but only 5,457 in pornographic websites. However, when looking specifically at ATS services, we see that they are more widespread and diverse in pornographic websites than in regular ones: 12% and 1% of all the third-party domains found in pornographic and regular websites are associated with ATSes, respectively. The intersection between the set of ATSes operating in the regular and pornographic websites is also low: only 86 third-party advertising and tracking services are present in both types of websites. This analysis reveals that a majority of advertising and tracking services operating in the online pornography ecosystem are unlikely to be present in regular websites. For instance, `exosrv.com` and `exoclick.com`, both belonging to Exoclick, are present in 2,709 pornographic websites (43% of the corpus) but only in 6 regular websites. These figures only represent a lower-bound estimation of the presence of advertising and tracking services in pornographic websites due to the well-known limitations of existing domain classifiers and blacklists [44, 71]. In Section 5, we will study the behavior of each third-party service to identify more trackers.

4.2.1 A closer look at the long-tail. The set of third-party services present in pornographic websites varies with the

[7] In some cases, the `Subject` field only contains the domain name of the website instead of the company name. We choose not to take the certificate information of these websites into account.
[8] Oracle operates a data marketplace, the largest third-party data marketplace for "open and transparent audience data trading" according to their own sources [64].

Table 3: Third-party presence by popularity interval (per Alexa's 2018 highest rank). For each interval we show the total number of third-party domains ("Total") and the third-party domains found only in this interval ("Unique")

Popularity Interval	Number of porn websites	Third-party domains (Unique to the interval)
0 — 1k	73	407 (119)
1k — 10k	536	1,327 (531)
10k — 100k	3,668	3,702 (2,115)
100k+	2,056	2,363 (1,007)

popularity of the hosting site. In other words, the more unpopular the pornographic website is, the more obfuscated and opaque are the third-party domains it embeds.

Table 3 shows the presence of third-party services in porn websites when grouped in different popularity intervals (according to their highest Alexa rank throughout 2018). Only 3% of third-party domains, regardless of their purpose, are present in all the different tiers of popularity. Amongst those we find cloud providers such as `cloudflare.com` and large advertising companies (*e.g.,* doubleclick by Alphabet), but also ATS companies specialized in adult sites such as `doublepimp.com` or `exoclick.com`. We would like to stress that Alphabet Inc. has specific policies about the type of content that can be distributed through their ad network as well as on the hosting site [35]. [9]

In order to get a better understanding of the implications of low popularity – and possibly reputation – in terms of third-party services embedded in porn website, we take a deeper look at 2,056 unpopular pornographic sites that never reach an Alexa ranking over 100K throughout 2018. This detailed analysis confirms that it is more likely to find advertisement and analytics services that are not commonly used by the prominent websites in unpopular pornographic websites. In fact, we identify that 18% of the third-party services embedded on porn websites, appear only in the less popular ones according to Alexa. Among others, we can find obscure analytic services like adultforce.com [7], for which we could not find a privacy policy on their homepage. We also found potentially malicious domains (according to Dr. Web) such as the traffic trade webpage `itraffictrade.com` [21].

4.2.2 An organization-level analysis. We now present an organizational-level analysis of the third-party domains operating in pornographic websites, regardless of their role. Figure 3 shows the 15 companies offering third-party services to most of the studied pornographic websites. As we

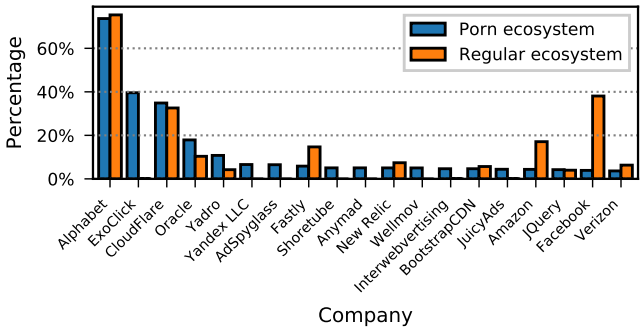

Figure 3: Most relevant third-party organizations in the porn ecosystem. We show their prevalence in the regular ecosystem for comparison.

can see, Alphabet is —as in the regular web— the most prevalent organization (74% of the total pornographic websites). Exoclick and Cloudflare services [10] are second and third with 40% and 35% of prevalence, respectively. When comparing with the third-party companies present in the regular web, we also find that several ones solely operate in the adult industry. While some of them are well-known actors like Exoclick [30], others are lesser known companies like JuicyAds (4%) [45] and EroAdvertising (5%) [26]. A similar case is Chaturbate, which, apart from being a porn website inside of our corpus, it also acts as a third-party service in 2% of the websites, acting as an ATS in 69% of the websites where it is present.

We find that the overall presence of Alphabet services (*e.g.,* Doubleclick, Google Analytics) is very similar in both regular and pornographic websites. These services appear in 74% of the porn websites but their presence varies greatly when considering individual services: `google-analytics.com` is present in 39% of porn websites, while `doubleclick.net` —an ad-network— appears in 12% of them (for reference, 60% of non-porn websites connect to Doubleclick domains). The higher presence of Oracle in porn websites is caused by its `addthis.com` service, which provides features like social network integration and content sharing (*e.g.,* pictures or videos). Another interesting case is the domain `alexa.com` which is related to the Amazon-owned browser extension that populates such list. We see that while Facebook is highly popular in the web ecosystem, its presence is greatly reduced in porn webpages.

Finally, we find the presence of `rlcdn.com`, a service that belongs to RalpLeaf which is TowerData/Acxiom [1], one of the largest data brokers in the world. Previous research has reported that Acxiom sells data to Facebook [13]. RalpLeaf

[9] Performing an analysis on whether the host sites are not complying with Google Ads Policies is outside the scope of this paper.

[10] In this specific case, we cannot confidently confirm that Cloudflare is operating these domains. It might be possible that other companies, advertising services, and trackers might be using Cloudflare's infrastructure.

service appears on 4 porn websites, one of which is classified as *bestiality*, an illegal sexual practice on countries like Switzerland among others.

5 PRIVACY RISKS

The (*i*) sensitive nature of pornographic content; and (*ii*) the quite unique ecosystem of third-party ATSes operating in them highlight the importance of studying in depth the behavior of these services and their use of tracking technologies. In this section, we perform a multi-dimensional analysis of the various privacy risks that visitors of pornographic websites might be exposed to (Section 5.1). We also provide an analysis of the use of insecure protocols (*e.g.,* HTTP) who may allow in-path observers like censors to monitor users' browsing habits (Section 5.2), and report on the anecdotal presence of malware in these sites (Section 5.3).

5.1 User Tracking Techniques

We leverage our customized version of OpenWPM to identify the use of various tracking techniques in pornographic websites, specifically HTTP cookies, cookie sync-ing, and advanced fingerprinting techniques.

5.1.1 HTTP Cookies. Online companies often use HTTP cookies as a mean for tracking users across the web. They do so by generating and storing unique identifiers in end-users' browsers. We use OpenWPM we identify 89,009 HTTP cookies installed by 92% different porn websites. This includes both first- and third-party service cookies. However, not all cookies might be used for the purpose of tracking users (*e.g.,* session cookies). Therefore, we focus our analysis in those HTTP cookies that may potentially contain user identifiers. Therefore, we discard session cookies and those with a length below 6 characters— unlikely to contain unique identifiers [33]. After applying this filter, we keep 51,648 HTTP cookies that can potentially be used to track users. 3% of them are larger than 1,000 characters, even reaching 3,600 characters in the case of cookies installed by by third-party ATS services like `juicyads.com`, `tsyndicate.com`, `exoclick.com`, and `exosrv.com` and other porn websites.

We put our focus on the 30,247 HTTP cookies installed by 3,343 third-party domains. These domains are responsible for cookies present in 72% of the porn websites. The 100 most popular cookies (by their unique *name = value* combination) actually appear in over 30% of the total porn websites. Moreover, as shown in Table 4, the main third-party services responsible for installing HTTP cookies in users' browsers are ExoClick, Oracle (AddThis), Yandex, and JuicyAds. While ExoClick and JuicyAds are specific to the online porn ecosystem, AddThis and Yandex are commonly found in regular web services, allowing these firms to potentially track users across both ecosystems.

Table 4: The 5 most common third-party domains delivering cookies that potentially contain unique IDs.

Third-party domain	% porn websites	# Cookies	ATS	In web ecosystem	% Cookies with user IP
`exosrv.com`	21%	2095	✓	✓	85%
`addthis.com`	17%	1289	✓	✓	0%
`exoclick.com`	14%	434	✓	✓	29%
`yandex.ru`	4%	312	✓	✓	0%
`juicyads.com`	4%	475	✓	✓	0%

Encoded Information in HTTP Cookies. We decode the cookie values using two types of codification, base64, and URL. We detect that there are 2,183 cookies storing the IP address of our physical machine and potential IDs. 97% of these cookies belong to Exoclick domains, which are found in 440 different porn websites. In Table 4, we can observe that most of the cookies that Exoclick and its subsidiary services deliver have embedded the user IP address in the cookie along with a potential ID. In particular, 85% of `exorsrv.com` cookies and 29% of `exoclick.com` cookies follow this pattern. Furthermore, we identify 28 cookies in 15 websites storing approximate geolocation data, potentially obtained through geo-IP databases [54]. 27 of these 28 cookies are delivered by two third-party domains, `fling.com` and `playwithme.com`. While the former only stores the coordinates, the latter also includes detailed information about the network infrastructure. The accuracy of geo-IP databases is not very precise in general, but it could reveal the precise location of a user in certain scenarios [66].

5.1.2 Cookie Synchronization. For security purposes, modern browsers limit the access to cookies to the service that has installed it. To circumvent this security mechanism and ease cross-site tracking, third-party services use a technique called cookie synchronization (cookie syncing in short).

Through this technique, online services can share their cookie data with others by embedding the cookie in the URL [36, 65]. We study the use and prevalence of cookie syncing in pornographic websites by checking if any of the observed HTTP cookies is later embedded in subsequent HTTP requests. To avoid introducing false positives, we don't split the cookie value by delimiters like "-" or "=" so our findings offer a lower bound estimation of the prevalence of this technique.

The number of pornographic websites for which we have observed this practice is 2,867. This covers 58% of the top-100 most popular porn websites according to Alexa. However, when matching the pairs of organizations (at the domain level) involved in this practice, we found 4,675 pairs as shown in Figure 4. This involves 1,120 origin and 727 destination services. Cookie syncing can also occur between domains belonging to the same organization. For instance, the

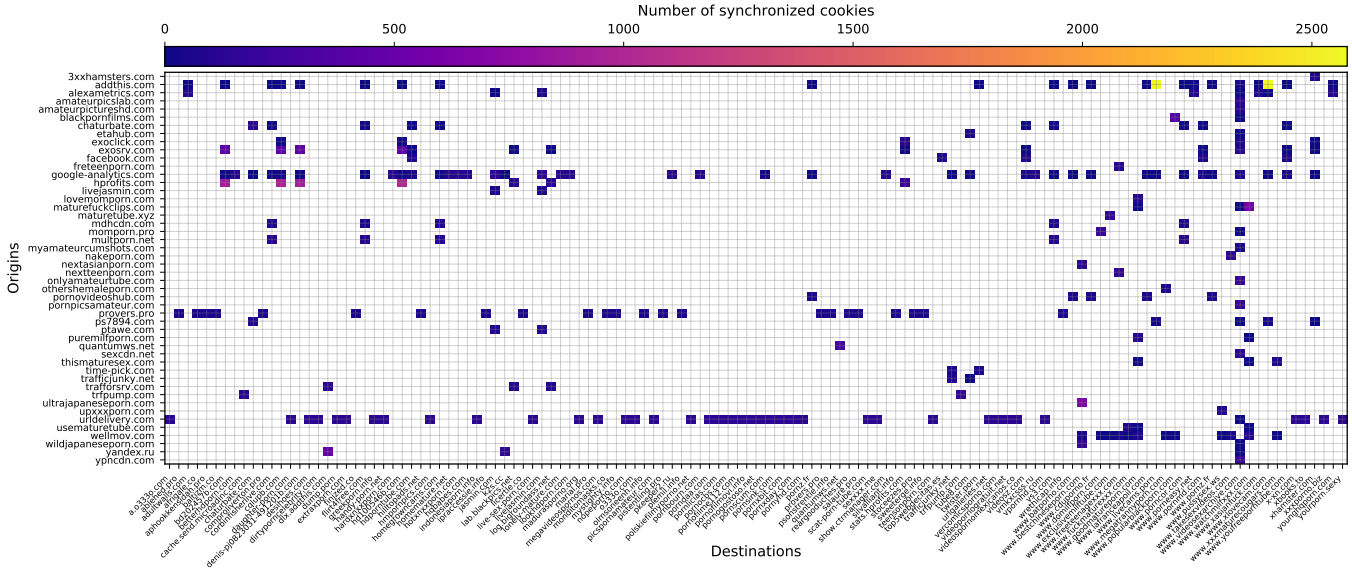

Figure 4: Cookie syncing between organizations. Pairs of domains that exchanged at least 50 cookies are shown.

Table 5: Third-party domains using different tracking-techniques. The `ATS` and `Regular web` columns indicate whether these services are indexed in EasyList/EasyPrivacy or if they are present in the regular web, respectively.

Domain	Presence in porn sites	ATS	Regular web	JavaScripts		
				Total	Canvas fingerprinting	WebRTC
adsco.re	152	-	✓	1	0	1
ero-advertising.com	33	✓	✓	32	32	0
cloudfront.net	31	✓	✓	8	8	0
cloudflare.com	28	✓	✓	2	2	0
adnium.com	26	✓	-	41	41	0
highwebmedia.com	22	✓	✓	1	1	0
xcvgdf.party	18	-	-	18	18	0
provers.pro	15	✓	-	1	1	0
montwam.top	13	✓	-	25	25	0
dditscdn.com	10	✓	✓	1	1	0

third-party domains `hd100546b.com` and `bd202457b.com` synchronize HTTP cookies with `hprofits.com`, among many other services. The X.509 certificate for these three domains suggest that all of them belong to `hprofits.com`, an ad exchange platform according to their website.

5.1.3 User Fingerprinting. Fingerprinting techniques allow trackers and services to create a unique user identifier by accessing and processing several characteristics of the user's device using JavaScript APIs. As opposed to cookie-based tracking, this sophisticated method can be used to persistently track users and their activities across websites without having to rely on cookies.

We focus our analysis in detecting services using either canvas or canvas font fingerprinting techniques [24]. HTML Canvas Fingerprinting is a tracking technique that exploits system differences between devices in how they render images. These scripts use the `CanvasRenderingContext2D` and the `HTMLCanvasElement` JavaScript APIs to generate images using specific height, width, different fonts, and background colors, among other characteristics. Font fingerprinting is a variation of canvas fingerprinting in which a tracker can leverage the fonts that each browser has installed to generate a unique ID of the device. This is achieved with the `measureText` method of the HTML Canvas API which allows to draw text using different fonts. Depending on the size of the written text, the service infers if the browser has a particular font. However, in both cases, not all the services that invoke these JavaScript APIs do so for the purpose of tracking users.To eliminate false positives, we follow the methodology proposed by Englehardt *et al.* [24]. In the case of canvas fingerprinting, we exclude: *i*) all the canvas with width and height below 16px; *ii*) we remove scripts that do not use at least two colors or text with more than 10 different characters; *iii*) those scripts that don't call either the `toDataURL` or the `getImageData` methods with an area below 320px; and *iv*) scripts that use the save, restore, or `addEventListener` methods of the rendering context. Despite these precautions, none of the scripts reported by OpenWPM meet these criteria. For font fingerprinting, we only consider those scripts that set the font property and call the `measureText` method on the same text at least 50 times. We identify 245 different JavaScripts performing canvas fingerprinting in 315 porn websites. 74% of them are delivered by 49 third-party services including `ero-advertising.com` and `highwebmedia.com`, a service that belongs to `chaturbate.com` (one of the biggest

sex live services). These third-party services are present in 4% of all the porn websites. We only detect one script delivered by online-metrix.net leveraging font fingerprinting. We find that the script performing font fingerprinting and the 91% of the scripts using canvas fingerprint were not previously indexed by tracking blacklists like EasyList and EasyPrivacy.

5.1.4 Other Potential Tracking Methods. We have identified instances of other methods that could be potentially used for tracking purposes. However, we did not gather sufficient evidence to demonstrate that these APIs are actually used for such purposes. One case is WebRTC [39], a technology to establish real-time peer-to-peer communications between browsers. WebRTC APIs allow collecting the IP address of the users, as well as the local network address. Through the combination of WebRTC with other tracking techniques [24], online services can discover networking information such as devices hosted behind the same NAT for cross-device tracking [73], or to identify whether the user connects through a VPN [46]. We have identified 27 different scripts using WebRTC, 21 of which use other tracking mechanisms in conjunction. The services using WebRTC along with other tracking technologies are present in 177 different porn websites. Two of the 13 different third-party services using WebRTC, appear in the regular web and are classified as ATS by EasyList. These services are traffichunt.com and online-metrix.net, an advertisement platform and a web analytic service, respectively.

5.2 (Lack of) Network Security Standards

We check for HTTPS support in porn websites using our OpenWPM crawler. By default, we crawl each website using HTTPS, only downgrading to HTTP when HTTPS is not supported by the server. Safeguarding users' privacy and security should be a priority for providers of pornographic content, particularly if users can be subject to censorship and surveillance at the network level [70, 92]. The use of encryption for transmitting data over the network is also a provision in privacy laws like the GDPR (Article 32 [18]) and CCPA [16].

Table 6 shows the use of HTTPS in pornographic websites depending of their highest Alexa rank in 2018. We find that over 97% of the most popular websites (in the top-1K of the Alexa ranking) do support HTTPS. However, the ratio of porn websites supporting HTTPS drops as their popularity does: HTTPS support decays to less than 30% for websites whose highest Alexa rank in 2018 was 10,000 or lower. This trend is similar for third-party services: those included in popular porn websites are more likely to support HTTPS. We find that 3,445 pornographic websites (54% of the corpus) are not fully HTTPS: either the website or one of its embedded third-party do not support HTTPS. Furthermore, 8% of these websites sending traffic in the clear embed cookies that we were able to decode in Section 5.1.1.

Table 6: HTTPS usage in pornographic websites

Interval	Feature	HTTPS
0 — 1k	Porn websites (30)	97%
	3rd-party services (107)	93%
1k — 10k	Porn websites (282)	71%
	3rd-party services (655)	60%
10k — 100k	Porn websites (758)	54%
	3rd-party services (1,022)	52%
100k+	Porn websites (5,754)	27%
	3rd-party services (4,473)	34%

5.3 Potential Malicious Behaviors

We conclude this section with a short study of the presence of potentially malicious behaviors in pornographic websites according to VirusTotal [89]. To minimize false positives, we only report domains flagged as malicious by at least 4 of the 70 different malware scanners aggregated by VirusTotal. There are 7 porn websites classified as a potentially malicious websites. Further, the potentially malicious activities also extend to 16 embedded third-party services in 41 porn websites. We highlight the presence of "coinhive.com", a cryptocurrency mining service present in 3 porn websites. This suggests that owners of pornographic websites also explore alternative monetization schemes beyond online advertisement and subscription-based models.

6 GEOGRAPHICAL DIFFERENCES

In this section, we investigate whether pornographic websites adapt their behavior to meet the requirements of different regulatory frameworks. For that, we launch our crawls from different vantage points using commercial VPNs and our Madrid-based crawler.

6.1 Third Party Services

Table 7 shows the number of third-party services embedded on porn websites per country. We can see that the total number of third-parties on each location remains stable but for Russia, which has over 700 third-party services less. However, when looking at individual instances of third-party services, we find that there are hundreds of domains that are unique in each country —around 10% of them are related to CDNs or porn websites that generate arbitrary domains like img100-589.xvideos.com. When analyzing the domains by their parent companies, we can see that

Table 7: Comparison of the domains found on porn ecosystem from different geographical points. The values do not include domains loaded dynamically on the websites.

	FQDN	Web Ecosystem	Unique Country	ATS	Unique Country
USA	5,483	16%	357	635	25
UK	5,364	15%	231	620	20
Spain	5,494	16%	561	592	59
Russia	4,750	16%	373	542	27
India	5,340	15%	275	607	21
Singapore	5,310	15%	233	608	16
Total	7,813	14%	2,030	816	168

most of the unique ATS per country are still Google services (*e.g.,* `adservice.google.ru` or `google.co.in`). Anecdotally, there is third-party service, `eventraliaclinental.info` that is present in 16 porn websites only from Russia and for which we could not find public information.

6.2 Malware Presence

We analyze the third-party services obtained globally using VirusTotal to detect potentially malicious activities. The number of services considered as potential malware in each country varies, from 15 services in Russia to 19 services in India. However, we observe that there are 13 services that are present regardless of users' geolocation: (*e.g.,* the crypto mining domain `coinhive.com`). The number of websites that contain potential malware domains varies from 29 websites in Russia to 42 in Spain. Furthermore 26 websites always contain potential malware regardless of the country of access.

7 REGULATORY COMPLIANCE

In this section, we evaluate pornographic website's regulatory compliance. Specifically: (1) we verify the presence and use of cookie consent forms as provisioned by the EU GDPR and ePrivacy regulation; and (2) the use of verifiable age verification mechanisms in the context of the UK's Digital Economy Act. We also investigate the lack of privacy policies and potential inconsistencies between these legal documents and the behavior observed in the pornographic websites.

7.1 Cookie Consent Notice

The ePrivacy directive will require websites to obtain consent from European users before installing and using cookies, unless the cookie is strictly necessary for the webpage functionality. As this legislation is not yet into effect, the use of cookies is currently regulated by the GDPR, which indicates that users must consent to the use of any technique that may uniquely identify them [18]. This is typically done through cookie consent forms.

Table 8: Usage of HTTP cookie banners in porn websites.

Type	EU	USA
No Option	1.36%	1.39%
Confirmation	2.82%	2.3%
Binary	0.2%	0.06%
Checkbox	0.03%	0.01%
Total (N = 6,843)	4.41%	3.76%

Degeling *et al.* performed a preliminary analysis of cookie consent forms (cookie banners) in the web after GDPR [19], analyzing 6,579 websites and finding that around 62% of them display a cookie consent-banner after GDPR. They presented a categorization of HTTP cookie banners into 6 different groups, that we will follow: *(i)* ***No Option****:* This type of cookie banner only informs users abut the use of HTTP cookies without giving the possibility of accepting or rejecting them; *(ii)* ***Confirmation****:* This type informs users about the use of cookies, but users can only show their accordance with the use of cookies, they can not reject them; *(iii)* ***Binary****:* In this case, users can accept or reject the use of cookies; *(iv)* ***Slider****:* This type of cookie banner gives users more fine-grained control over the level and type of cookies, that they allow by adjusting a slider; *(v)* ***Checkbox****:* This type of banner gives users the capacity to allow/reject cookies for a specific purpose or from a particular third-party service; *(vi)* ***Other****:* Any other type of banner that does not match any of the above. These banners tend to have a higher degree of complexity.

Identifying cookie banners automatically in websites is not trivial and due to technical limitations, we could only instrument our customized OpenWPM to identify the following types: No option, Confirmation, Binary and Checkbox. We do so by identifying whether there's an HTML element in the DOM containing a banner. If so, we extract the text, taking also a screenshot of it for manual inspection and validation. We perform the crawling of webpages from Spain and from the USA, in this latter case using a VPN provider, in order to study both GDPR and CCPA compliance. Table 8 shows the percentage of pornographic websites in which we found HTTP cookie banners. As can be observed, the proportion of pornographic webpages with cookie banners is very small, being only 4% of the total. A second observation from Table 8 is that the difference between accessing webpages from one country or the other is also very small, as only 0.6% more pages show a cookie banner when accessing from Europe.

The low presence of cookie banners is remarkable when compared with the fact that 72% of the total pornographic

websites contain third-party cookies (Section 5.1.1). Moreover, out of the websites that show a cookie banner, 32% do not give users any control over the use of cookies as the banner only shows information. While possibly not all third-party cookies are used for tracking purposes, these numbers suggest that many websites may be in a potential violation of the GDPR, particularly given the sensitive nature of pornographic websites. In comparison, Degeling *et al.* showed that 69.9% out of a corpus of 6,357 websites had a cookie consent banner in January 2018 [19]. It is important to note that our methodology uses OpenWPM to crawl the websites and that we do not interact with the webpage once we have visited it. Therefore, even in those websites where a cookie banner is present, we never gave consent to the use of cookies. Section 5.1.1 shows that we still found cookie presence in porn websites regardless of not granting permission. As a final note, we have not found any clear correlation between the use of cookie consent forms and the popularity of porn webpages.

7.2 Age Verification

Some pornographic websites have taken positive steps to implement age verification mechanisms in an effort to comply with increasing regulatory pressures (see Section 2.1). In this section, we study how prevalent and how effective age-verification mechanisms are in the wild. For that, we use our Selenium-based crawler to parse the landing page of each porn website, and look for warnings and consent forms displayed to the user (Section 3.1). As our approach relies on string matching to identify such warnings, it is prone to introduce false positives, specially so in age-related keywords that appear often in the content of the websites. Therefore, we investigate a subset of the top-50 most popular pornographic websites manually.

We perform this manual analysis in 4 countries (the US, the UK, a second EU country, and Russia) to identify regional differences. The results obtained when accessing these websites from the USA, UK and the EU are exactly the same. The same set of 20% pornographic websites implement the same age verification mechanism, consisting of a simple warning text and a button to be clicked on. Instead, we have found significant differences when accessing these websites from Russia: only 14% of the analyzed websites have an age verification mechanism. Additionally, 8% of websites that do not verify users' age for the rest of countries do so in Russia whereas 12% of websites do not verify user age in Russia contrary to the rest of countries studied.

Despite regulatory pressures, the age verification mechanisms implemented by all these sites are easy to bypass so they could not be considered as "verifiable age verification mechanisms". If our automatic crawler manages to bypass the mechanism, a child would do it as well. We only found one webpage in Russia, `pornhub.com`, implementing a complex age verification mechanisms through social media accounts as requested by the federal government in 2017 [88].

7.3 Privacy Policies vs. Reality

The GDPR [18] requires all websites collecting or processing personal identifiable data from European citizens to portray a privacy policy describing their practices. The GDPR also requires that when a website collects private data, their privacy policy must clearly inform the user about such practices, including data collected by embedded third parties. We perform a best-effort crawl to collect privacy policies, if available, of each pornographic website to crosscheck with our empirical results, and highlight potential privacy violations. We perform this analysis using the method introduced in Section 3.1, only from a physical machine hosted in Spain.

Only 16% of the pornographic websites in our dataset have privacy policies reachable from the main page. We get these numbers after a manual sanitization of our results in which we check the shortest privacy policies found and remove 44 false positives caused by HTTP errors (response codes). We find that 218 (20%) of these privacy policies make an explicit mention to the GDPR. We dive deeper into the analysis of the privacy policies by first looking at length patterns. On average, privacy policies contain 17,159 letters with the shortest policy having 1,088 and the largest 243,649.

We also take a look at the similarity of the text in privacy policies for each website. We use the term frequency–inverse document frequency (TF-IDF) [74] to measure the similarity between two texts.[11] 76% of the 1,202,312 analyzed pairs have a similarity above 0.5. We find that this can be a direct result of websites from the same company having a very similar privacy policy as well as the prevalence of given templates that are highly popular across websites. The opacity of the privacy policies renders difficult performing an automatic analysis of their content at scale. To tackle this issue, we use the publicly available tool Polisis[37], which presents a summarized version of the privacy policy, to extract third-party entities and data collection methods. We do not have access to results in a machine-readable format so we rely on the web version of the tool to further investigate the top 25 websites tracking users (*i.e.,* Canvas fingerprinting and cookies) according to our results from Section 5. We manually asses that 72% of this subset of porn websites have a privacy policy in which they clearly state the use of cookies, the type of data collected, and the presence of third parties in their websites. Nevertheless, we find that only one of the

[11] The value goes from -1 (exactly opposite) to 1 (exactly equal) going through 0 (no co-relation).

websites disclose in their privacy policies their complete list of third-party advertising and tracking services.

These findings show that, while privacy policies are becoming more common, complete, and clear to users, there are still many websites with embedded third-party tracking services without privacy policies. When they do, with only one exception, they do not disclose the whole list of embedded third-parties.

8 RELATED WORK

The research community has studied web tracking extensively. We expand the state of the art by looking for the first time at a specific but highly sensitive ecosystem that, to this day, has remained unexplored. We discuss below studies that are more relevant to our work.

Web tracking: Several research studies have made groundbreaking contributions to illuminate and uncover the privacy risks of the web. This includes studying the use of HTTP cookies, cookie syncing [5, 24, 25, 47, 65, 75], persistent tracking mechanisms [5, 50, 85], and advanced fingerprinting techniques [24, 32, 57]. In [12], Bashir *et al.* introduced the notion of inclusion chain to model the diffusion of user's data within third parties. Zimmeck *et al.* studied the prevalence of cross-device tracking and the implications for user privacy [95]. These studies were possible thanks to web crawlers: either customized versions of Chrome or Firefox [50, 58], or headless browsers such as Selenium [69, 77, 90] or purpose-specific ones like OpenWPM [23, 24, 33, 56].

Pornographic websites: The online porn industry has remained largely underground. There has been isolated steps towards studying this ecosystem, mainly from a content availability standpoint in a major porn website [83]. Vasey and Abild tackle the topic of pornography in the Internet[86], comparing what people say about their sexuality with the results of billions of Internet searches. Wondracek *et al.* studied the economic structure of the online porn industry [91], showing that these websites usually present shady schemes to generate revenue (such as traffic trading). Altaweel *et al.* used OpenWPM to study user tracking in 11 of the most popular pornographic websites [9]. They showed that there is a lesser presence of tracking in porn websites in comparison with popular non-pornographic sites. Marotta-Wurgler studied the presence of privacy policies in websites [53], including 17 popular adult websites showing that this type of website is more likely to include privacy policies. These differences with our findings can be explained with the fact that our corpus of websites is significantly bigger and it includes less popular websites.

Regulatory compliance: We build on previous work to study GDPR compliance and measure websites' readiness for the ePrivacy directive and UK's Digital Economy Act age-verification. Degeling *et al.* studied the impact of GDPR on web privacy and the prevalence of cookie consent mechanisms in a corpus of 6,579 websites in each of the 28 EU member states [19]. Other studies have looked at the presence of privacy policies across websites [94] and mobile apps [62, 78] and ways to automatically study such policies [37]. Kulyk *et al.* performed a user study on user's reaction to cookie consent notices in 50 German websites [48]. Finally, Marotta-Wurgler [53] found that most of the porn websites have policies. However, they only looked at the 17 most relevant porn websites.

9 FUTURE WORK

Our study has opened a number of research questions that we plan to address in the near future.

Our study focuses in basic aspects of GDPR compliance and UK's efforts to control minors' access to pornographic content. This analysis could be extended to look deeper into GDPR compliance, for instance by further analyzing the values of cookies to investigate the prevalence of other tracking IDs. An interesting aspect to study in the future could be characterizing cross border data exchanges as reported by Iordanou *et al.* [43] and Razaghpanah *et al.* [71], or performing a deeper investigation of the connections between online trackers, advertising services, and data brokers. Another aspect that could be analyzed in future work are the privacy implications on the subscription websites, analyzing which type of data they required to create the account as well as compare the presence and amount of tracking services between the subscription websites and their free mode in case that they have it.

Finally, in this study, we have intentionally not studied the performance of anti-tracking technologies to protect users' privacy, including safe-browsing modes and popular ad-blockers. We believe that analyzing the effectiveness of such tools in specific ecosystems and longitudinally deserves a dedicated study on its own.

10 CONCLUSIONS

Online porn has been traditionally considered as an obscure subsystem of the Internet. In this paper we performed the first comprehensive and large scale analysis targeted at analysing the privacy risks and (lack of) regulatory compliance of porn sites. The porn industry is not different in many aspects from regular web services: it has rapidly integrated advanced tracking technologies to monitor (and in some cases to monetize) users. Yet there are noticeable differences with regards to regular websites, specially regarding the parallel ecosystem of third parties offering advertising and tracking services to online porn websites.

The presence of porn-specific trackers might render many anti-tracking technologies based on blacklists insufficient. We find that the 91% of the scripts implementing canvas fingerprinting were not indexed by EasyList and EasyPrivacy lists. Furthermore, we demonstrate that a large number of porn websites fail to implement the most common security mechanisms, such as the use of HTTPS and basic transparency requirements such as privacy policies and cookie consent forms; even in those websites actively tracking users. Only the companies behind some of the most popular pornographic websites seem to make positive efforts to comply with current legislation, possibly fearing the high fines of new regulations like the GDPR. Besides data protection and users' privacy aspects, we demonstrated that the efforts made by the online porn industry to prevent children access to inappropriate content are not being widely deployed. While most countries do not have laws to prevent children's' access to pornographic material, we have observed that the deployment of these mechanisms is rare even in jurisdictions where such laws will be applicable soon.

We believe that our work opens new doors for other studies in semi-decoupled web subsystems offering sensitive services that are popular yet not widely studies (*e.g.,* gambling and online health services), while also informing the public debate. Many of these services might fall between the cracks of public scrutiny due to being specialized in a unique ecosystem.

ACKNOWLEDGEMENTS

This work is partially supported by the Spanish grant TIN2017-88749-R (DiscoEdge), the Region of Madrid EdgeData-CM program (P2018/TCS-4499), the European Union's Horizon 2020 Innovation Action program (grant Agreement No. 786741, SMOOTH Project) and "la Caixa" Foundation agreement LCF/PR/MIT17/11820009.

REFERENCES

[1] Acxiom. https://www.acxiom.com.
[2] Only4 adults. only4adults.com.
[3] Top porn sites. toppornsites.com.
[4] Alexa Websites Ranking, 2019. https://www.alexa.com/topsites/.
[5] Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., and Diaz, C. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 674–689.
[6] Addthis. About AddThis. https://www.addthis.com/about/oracle/.
[7] Adult Force. Homepage. https://www.adultforce.com/#/.
[8] AgeID. AgeID announced to the industry at European Summit. https://www.ageid.com/press/article/11.
[9] Altaweel, I., Hils, M., and Hoofnagle, C. J. Privacy on adult websites.
[10] ASACP. ASACP Members. https://www.asacp.org/index.php?content=members#top.
[11] ASACP. Association of Sites Advocating Child Protection. https://www.asacp.org/.
[12] Bashir, M. A., and Wilson, C. Diffusion of user tracking data in the online advertising ecosystem. *Proceedings on Privacy Enhancing Technologies 2018*, 4 (2018), 85–103.
[13] BBC. Facebook scandal: Who is selling your personal data? https://www.bbc.com/news/technology-44793247.
[14] BBC News. Russia extends porn site ban. https://www.bbc.com/news/technology-37373244.
[15] Bluekai. Oracle Buys Bluekai. https://www.oracle.com/es/corporate/acquisitions/bluekai/.
[16] California State Legislature. California Consumer Privacy Act. https://www.caprivacy.org/.
[17] CookieBot. The EU ePrivacy Regulation and Cookies - What do I need to do? https://www.cookiebot.com/en/eprivacy-regulation-and-cookies/.
[18] Council of European Union. General Data Protection Regulation 679/2016, 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN.
[19] Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096* (2018).
[20] Disconnect Tracking Protection List. https://github.com/disconnectme/disconnect-tracking-protection.
[21] Dr. Web. Homepage. https://www.drweb.com/.
[22] EasyList. EasyList. https://easylist.to.
[23] Englehardt, S., Eubank, C., Zimmerman, P., Reisman, D., and Narayanan, A. Openwpm: An automated platform for web privacy measurement. *Manuscript* (2015).
[24] Englehardt, S., and Narayanan, A. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016), ACM, pp. 1388–1401.
[25] Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., and Felten, E. W. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web* (2015), International World Wide Web Conferences Steering Committee, pp. 289–299.
[26] EroAdvertising. Homepage. https://www.eroadvertising.com/#!/.
[27] EUR-Lex. ePrivacy proposal, 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017PC0010.
[28] European Comission. The EU Internet handbook. http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm.
[29] European Comission (Algolia). GDPR Article 9. https://gdpr.algolia.com/gdpr-article-9.
[30] ExoClick. Homepage. https://www.exoclick.com/.
[31] Falahrastegar, M., Haddadi, H., Uhlig, S., and Mortier, R. Tracking personal identifiers across the web. In *International Conference on Passive and Active Network Measurement* (2016), Springer, pp. 30–41.
[32] Fifield, D., and Egelman, S. Fingerprinting web users through font metrics. In *International Conference on Financial Cryptography and Data Security* (2015), Springer, pp. 107–124.
[33] Fouad, I., Bielova, N., Legout, A., and Sarafijanovic-Djukic, N. Tracking the pixels: Detecting web trackers via analyzing invisible pixels. *arXiv preprint arXiv:1812.01514* (2018).
[34] Futurism. Data From British Porn Viewers Might Be In The Hands of One Company, 2018. https://futurism.com/mindgeek-monopoly-uk-

porn-viewers-data.

[35] Google. Google Policies Help - Adult Content. https://support.google.com/adspolicy/answer/6023699?hl=en, 2019. Accessed: February 12, 2019.

[36] Google Developers. Cookie matching. https://developers.google.com/authorized-buyers/rtb/cookie-guide.

[37] Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., and Aberer, K. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)* (2018), pp. 531–548.

[38] IAB and PWC. The Official xHamster 2019 Trend Report, 2019. https://xhamster.com/blog/posts/911001.

[39] IAB and PWC. WebRTC, 2019. https://webrtc.org/.

[40] Indian Government. The Personal Data Protection Bill. https://meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf.

[41] International Amnesty. Uganda's new anti-human rights laws aren't just punishing LGBTI people. https://www.amnesty.org.uk/uganda-anti-homosexual-act-gay-law-free-speech.

[42] International Association of Privacy Profesionals. GDPR matchup: Japan's Act on the Protection of Personal Information. https://iapp.org/news/a/gdpr-matchup-japans-act-on-the-protection-of-personal-information/.

[43] Iordanou, C., Smaragdakis, G., Poese, I., and Laoutaris, N. Tracing cross border web tracking. In *Proceedings of the Internet Measurement Conference 2018* (2018), ACM, pp. 329–342.

[44] Iqbal, U., Shafiq, Z., Snyder, P., Zhu, S., Qian, Z., and Livshits, B. Adgraph: A machine learning approach to automatic and effective adblocking. *arXiv preprint arXiv:1805.09155* (2018).

[45] Juicy Ads. Homepage. https://www.juicyads.com/.

[46] Khan, M. T., DeBlasio, J., Voelker, G. M., Snoeren, A. C., Kanich, C., and Vallina-Rodriguez, N. An empirical analysis of the commercial vpn ecosystem. In *Proceedings of the Internet Measurement Conference 2018* (2018), ACM, pp. 443–456.

[47] Krishnamurthy, B., and Wills, C. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web* (2009), ACM, pp. 541–550.

[48] Kulyk, O., Hilt, A., Gerber, N., and Volkamer, M. "this website uses cookies": Users' perceptions and reactions to the cookie disclaimer,". In *European Workshop on Usable Security (EuroUSEC)* (2018).

[49] Legislation.gov.uk. Digital Economy Act 2017. http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted.

[50] Lerner, A., Simpson, A. K., Kohno, T., and Roesner, F. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *25th {USENIX} Security Symposium ({USENIX} Security 16)* (2016).

[51] Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* (1966).

[52] Luxembourg times. Porn empire reports half billion dollars in revenue – but ends year with loss, 2018. https://luxtimes.lu/luxembourg/33248-porn-empire-reports-half-billion-dollars-in-revenue-but-ends-year-with-loss.

[53] Marotta-Wurgler, F. Self-regulation and competition in privacy policies. *The Journal of Legal Studies 45*, S2 (2016), S13–S39.

[54] MaxMind. Detect online fraud and locate online visitors. https://www.maxmind.com/en/home.

[55] McDonald, A., Bernhard, M., Valenta, L., VanderSloot, B., Scott, W., Sullivan, N., Halderman, J. A., and Ensafi, R. 403 forbidden: A global view of cdn geoblocking. In *Proceedings of the Internet Measurement Conference 2018* (2018), ACM, pp. 218–230.

[56] Miramirkhani, N., Starov, O., and Nikiforakis, N. Dial one for scam: A large-scale analysis of technical support scams. *arXiv preprint arXiv:1607.06891* (2016).

[57] Mowery, K., and Shacham, H. Pixel perfect: Fingerprinting canvas in html5. *Proceedings of W2SP* (2012), 1–12.

[58] Mughees, M. H., Qian, Z., and Shafiq, Z. Detecting anti ad-blockers in the wild. *Proceedings on Privacy Enhancing Technologies 2017*, 3 (2017), 130–146.

[59] mypornbible. My porn bible. mypornbible.com.

[60] New York Magazine. The Geek-Kings of Smut. http://nymag.com/news/features/70985/index4.html#.

[61] NordVPN, 2019. https://nordvpn.com.

[62] Okoyomon, E., Samarin, N., Wijesekera, P., Elazari Bar On, A., Vallina-Rodriguez, N., Reyes, I., Feal, Á., and Egelman, S. On the ridiculousness of notice and consent: Contradictions in app privacy policies. In *Workshop on Technology and Consumer Protection* (2019), ConPro '19.

[63] OpenNet Initiative. Iraq. https://opennet.net/research/profiles/iraq#footnote24_is5a386.

[64] Oracle. Oracle Data Marketplace. https://docs.oracle.com/en/cloud/saas/data-cloud/data-cloud-help-center/Help/AudienceDataMarketplace/AudienceDataMarketplace.html.

[65] Papadopoulos, P., Kourtellis, N., and Markatos, E. P. Cookie synchronization: everything you always wanted to know but were afraid to ask. *arXiv preprint arXiv:1805.10505* (2018).

[66] Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. Ip geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review 41*, 2 (2011), 53–56.

[67] Pornhub. 2018 Year in Review, 2018. https://www.pornhub.com/insights/2018-year-in-review#2018.

[68] PrivateVPN, 2019. https://privatevpn.com/.

[69] Pujol, E., Hohlfeld, O., and Feldmann, A. Annoyed users: Ads and ad-block usage in the wild. In *Proceedings of the 2015 Internet Measurement Conference* (2015), ACM, pp. 93–106.

[70] Razaghpanah, A., Li, A., Filasto, A., Nithyanand, R., Ververis, V., Scott, W., and Gill, P. Exploring the design space of longitudinal censorship measurement platforms. *arXiv preprint arXiv:1606.01979* (2016).

[71] Razaghpanah, A., Nithyanand, R., Vallina-Rodriguez, N., Sundaresan, S., Allman, M., Kreibich, C., and Gill, P. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem.

[72] Reuters. Uganda's "kill the gays" bill shelved again. https://af.reuters.com/article/topNews/idAFJOE74C0HP20110513.

[73] Richter, P., Wohlfart, F., Vallina-Rodriguez, N., Allman, M., Bush, R., Feldmann, A., Kreibich, C., Weaver, N., and Paxson, V. A multi-perspective analysis of carrier-grade nat deployment. In *Proceedings of the 2016 Internet Measurement Conference* (2016), ACM, pp. 215–229.

[74] Roelleke, T., and Wang, J. Tf-idf uncovered: A study of theories and probabilities. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2008), SIGIR '08, ACM, pp. 435–442.

[75] Roesner, F., Kohno, T., and Wetherall, D. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (2012), USENIX Association, pp. 12–12.

[76] Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S. D., and Vallina-Rodriguez, N. A long way to the top: Significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference 2018* (New York, NY, USA, 2018), IMC '18, ACM, pp. 478–493.

[77] Selenium. What is Selenium? https://www.seleniumhq.org/.

[78] Story, P., Zimmeck, S., and Sadeh, N. Which apps have privacy policies? In *Annual Privacy Forum* (2018), Springer, pp. 3–23.

[79] The Guardian. Gay relationships are still criminalised in 72 countries, report finds, 2017. https://www.theguardian.com/world/2017/jul/27/gay-relationships-still-criminalised-countries-report.

[80] The Independent. Porn website age verification tool officially announced within UK, 2018. https://www.independent.co.uk/life-style/porn-age-verification-tool-uk-announcement-pornhub-ageid-adult-content-websites-mindgeek-a8242476.html.

[81] The Next Web. The (almost) invisible men and women behind the world's largest porn sites, 2016. https://thenextweb.com/insider/2016/03/03/the-almost-invisible-men-and-women-behind-the-worlds-largest-porn-sites/.

[82] Top websites. Adult catogory. Alexa. http://alexa.com/topsites/category/Top/Adult.

[83] Tyson, G., Elkhatib, Y., Sastry, N., and Uhlig, S. Demystifying porn 2.0: A look into a major adult video streaming website. In *Proceedings of the 2013 conference on Internet measurement conference* (2013), ACM, pp. 417–426.

[84] USA Today. China creates stern Internet, e-mail rules. https://usatoday30.usatoday.com/tech/news/2002/01/18/china-internet.htm.

[85] Vallina-Rodriguez, N., Sundaresan, S., Kreibich, C., and Paxson, V. Header enrichment or isp enrichment?: Emerging privacy threats in mobile networks. In *Proceedings of the 2015 ACM SIGCOMM Workshop on Hot Topics in Middleboxes and Network Function Virtualization* (2015), ACM, pp. 25–30.

[86] Vasey, P. L., and Abild, M. A billion wicked thoughts: What the internet tells us about sexual relationships, 2013.

[87] verification under the Digital Economy Act 2017, A. BBFC. https://www.ageverificationregulator.com/.

[88] Vice. Russians now need a passport to watch Pornhub. https://news.vice.com/en_us/article/kzgkv3/russians-now-need-a-passport-to-watch-pornhub.

[89] Virus Total. Virus Total. https://www.virustotal.com.

[90] Wang, L., Dyer, K. P., Akella, A., Ristenpart, T., and Shrimpton, T. Seeing through network-protocol obfuscation. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (2015), ACM, pp. 57–69.

[91] Wondracek, G., Holz, T., Platzer, C., Kirda, E., and Kruegel, C. Is the internet for porn? an insight into the online adult industry. In *WEIS* (2010).

[92] Yadav, T. K., Sinha, A., Gosain, D., Sharma, P. K., and Chakravarty, S. Where the light gets in: Analyzing web censorship mechanisms in india. In *Proceedings of the Internet Measurement Conference 2018* (2018), ACM, pp. 252–264.

[93] Yu, Z., Macbeth, S., Modi, K., and Pujol, J. M. Tracking the trackers. In *Proceedings of the 25th International Conference on World Wide Web* (2016), International World Wide Web Conferences Steering Committee, pp. 121–132.

[94] Zaeem, R. N., and Barber, K. S. A study of web privacy policies across industries. *J Info. Priv. Sec* (2017), 1–17.

[95] Zimmeck, S., Li, J. S., Kim, H., Bellovin, S. M., and Jebara, T. A privacy analysis of cross-device tracking. In *26th {USENIX} Security Symposium ({USENIX} Security 17)* (2017), pp. 1391–1408.

A ETHICAL CONSIDERATIONS

The data collection process does not involve human subjects. All the experiments were ran on a controlled environment using crawlers. The processes involving manual inspection were conducted by the authors of the paper. The members of the research team gave their approval to conduct such work, being aware of the possibility of having to see potentially uncomfortable images in some cases. Also, before running any experiments using the VPNs, we contacted NordVPN and PrivateVPN to inform them about our research work, in order to do not break their terms of uses and make sure that no harm to users were going to be caused.

Furthermore, we do not interact with the consent notices displayed by the websites and we do not surf beyond the landing page to avoid generating advertising revenues and accessing specific content. We do not discard the presence of additional tracking mechanisms and services beyond the landing page.

Before performing our data collection, we also defined a protocol to report any service distributing illegal pornographic content to the authorities in case that this uncomfortable situation arose. Unfortunately, we found one service distributing such content while performing our sanitization process. We immediately reported the case to the national authorities.