Privacy lost in online education: analysis of web tracking evolution

Su, Z.; Helles, R.; Al-Laith, A.; Veilahti, A.; Saxena, A.; Simonsen, J.G.; ... ; Cui, N.

Privacy Lost in Online Education: Analysis of Web Tracking Evolution

Zhan Su[1], Rasmus Helles[1(✉)], Ali Al-Laith[1], Antti Veilahti[1], Akrati Saxena[2], and Jakob Grue Simonsen[1]

[1] University of Copenhagen, Copenhagen, Denmark
{zhan.su,alal,simonsen}@di.ku.dk,
{rashel,antti}@hum.ku.dk
[2] Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands
a.saxena@liacs.leidenuniv.nl

Abstract. Digital tracking poses a significant and multifaceted threat to personal privacy and integrity. Tracking techniques, such as the use of cookies and scripts, are widespread on the World Wide Web and have become more pervasive in the past decade. This paper focuses on the historical analysis of tracking practices specifically on educational websites, which require particular attention due to their often mandatory usage by users, including young individuals who may not adequately assess privacy implications. The paper proposes a framework for comparing tracking activities on a specific domain of websites by contrasting a sample of these sites with a control group consisting of sites with comparable traffic levels, but without a specific functional purpose. This comparative analysis allows us to evaluate the distinctive evolution of tracking on educational platforms against a standard benchmark. Our findings reveal that although educational websites initially demonstrated lower levels of tracking, their growth rate from 2012 to 2021 has exceeded that of the control group, resulting in higher levels of tracking at present. Through our investigation into the expansion of various types of trackers, we suggest that the accelerated growth of tracking on educational websites is partly attributable to the increased use of interactive features, facilitated by third-party services that enable the collection of user data. The paper concludes by proposing ways in which web developers can safeguard their design choices to mitigate user exposure to tracking.

Keywords: Web-tracking · Information Security · Privacy · Online Education

1 Introduction

Privacy lost occurs when an individual's personal information or data is disclosed, shared, or accessed by others without their permission, which can result in various negative consequences, such as identity theft, financial fraud, damage to reputation, and discrimination. To investigate these issues, researchers may examine the historical practice of third-party web tracking, as described

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 440–455, 2023.
https://doi.org/10.1007/978-3-031-46664-9_30

by [17]. Third-party web tracking involves third-party entities, such as advertisers, social media widgets, and website analytics engines, that are embedded in the first-party sites that users directly visit and are capable of re-identifying users across domains while they browse the web. The proliferation of web tracking has spurred a growing body of research in the computer security and privacy community, which seeks to understand, quantify, and counteract these privacy risks posed by tracking companies compiling lists of websites that users have visited [2,3,17].

As the education industry transitions from traditional offline models to online or hybrid models, the need for privacy protection on educational websites is becoming increasingly prominent. This issue is crucial because the loss of privacy on educational websites can undermine the fundamental principles of privacy and security that are essential for individuals to feel safe and empowered while using the internet for educational purposes. By protecting users' privacy, educational websites can promote trust, openness, and responsibility, which are essential for fostering a positive and inclusive online learning experience. Therefore, several researchers have started studying the practice of web tracking in educational websites [11,12,21,24].

To deepen our comprehension of the nature and progression of tracking on educational websites, we propose an analytical framework that enables a comparative analysis of tracking on a specific type of site (in this case, education) in relation to a control group of sites with comparable traffic levels but of different types. The framework involves three steps: we construct a sample of educational websites, and a control sample of non-educational websites that have similar levels of traffic (Sect. 3.1). We then retrieve the historical websites from the Internet Archive's Wayback Machine[1] for both samples. Third, we scan the HTML file snapshots of the collected websites using the Wayback Machine (Sect. 3.2), and extract third-party trackers embedded in the HTML files (Sect. 3.3).

We aim to answer the following research questions, which we present along with our main findings:

RQ1: How has the use of trackers on educational websites evolved from 2012 to 2021?
In Sect. 4.1, we examine the average number of trackers from 2012 to 2021 and observe a general trend of tracker growth. Until 2018, both educational and non-educational sites sees substantial growth, but they diverge around the time of the introduction of the GDPR in 2018: at this point there is a minor drop in tracking on non-educational sites, which is not seen on educational sites, where the development merely stagnates.

RQ2: How does the evolution of the use of trackers differ between educational and non-educational websites?
Section 4.1 also addresses differences between educational and non-educational websites in the evolution of tracking between 2012 and 2021. The results show that despite the similarity of the underlying trend, the intensity of tracking has

[1] https://archive.org/.

grown relatively more on educational sites and that the growth has not similarly reverted as on non-educational websites after the introduction of the GDPR. The results are further supported by a Wilcoxon signed rank (WS) test conducted in Sect. 4.2, demonstrating that the intensity of tracking on educational sites surpassed that of non-educational sites in 2017.

RQ3: Is there a qualitative difference in what kind of trackers are used on educational and non-educational websites?
The quantitative difference between tracking on educational and non-educational sites that we find in the average number of trackers also shows up in the different compositions of the portfolios of trackers found at the two types of sites. We substantiate this statistically by using the Kolmogorov-Smirnov test (KS) to compare the distribution of trackers in these two groups of sites. To investigate the source of these differences, Sect. 4.3 examines the occurrence of some of the most popular trackers, demonstrating that the use of Twitter, Youtube, and Facebook has evolved very differently between educational and non-educational websites. In addition, Sect. 4.5 compares the presence of trackers presenting particular categories, demonstrating that tracking related to enhancing customer interaction in particular seems to have become relatively more common on educational websites over the past few years.

Our contributions can be summarized in two main points: (I) We develop a list of both educational and non-educational websites to investigate the issue of *privacy lost* in online education. The complete code and dataset we compiled can be accessed at[2]. (II) We conduct a quantitative and qualitative analysis of third-party tracking on educational websites, focusing on third-party services from 2012 to 2021. Our findings highlight potential concerns regarding the autonomy and fairness of education.

2 Related Work

Tracking through third-party cookies and scripts has been extensively studied from various perspectives. A significant portion of this research has focused on mapping the prevalence of trackers across samples of websites, such as those found on the Alexa top lists [1,9,19]. Other studies have investigated tracking on different platforms, including the mobile ecosystem [5,6,16].

Karaj et al. [13] proposed a method for measuring web tracking using a browser extension, resulting in a dataset covering 1.5 billion page loads collected over 12 months period from real users. Krishnamurthy and Wills [9] presented a dataset on tracking based on a crawl of the top 1 million websites. They developed an open-source web privacy measurement tool called OpenWPM, which allows researchers to detect, quantify, and characterize emerging online tracking behaviors. Our work is related to several general areas:

[2] https://github.com/shuishen112/Privacy_Lost.git.

Historical Web Tracking. Krishnamurthy and Wills provided early insights into web tracking, demonstrating the evolution of third-party organizations between 2005 and 2008 [15]. Lerner et al. presented longitudinal measurements of third-party web tracking behaviors from 1996–2016 [17]. Karaj et al. conducted a large-scale and long-term measurement of online tracking based on real users [13]. Agarwal and Sastr analyzed the top 100 Alexa websites over 25 years using data from the Internet Archive, studying changes in website popularity and examining different categories of websites and their popularity trends over time [2]. Amos et al. curated a dataset of 1,071,488 English language privacy policies spanning over two decades and encompassing more than 130,000 different websites [3].

Web Tracking after GDPR. Numerous studies have investigated web tracking following the implementation of the GDPR (General Data Protection Regulation) in the EU in May 2018, which imposed constraints on online data collection. These studies generally indicate a pattern of diminished tracking activity [7,20,22], but they also reveal that most sites appear unable or unwilling to fully comply with regulations [10,14,23], and tracking companies can still likely monitor user behavior [20].

Web Tracking in Educational Websites. A body of research focuses explicitly on educational websites, which are known to have a higher incidence of tracking technology than sites aimed at minors [24]. In particular, university websites exhibit a substantial prevalence of major tracking companies (e.g., Google, Facebook) [12]. While several recent papers discuss the implications of tracking on educational websites, there seems to be a lack of studies investigating third-party tracking on substantial samples of educational websites post-2018 or examining the development of tracking over time for these websites.

3 Data Collection

We provide a concise overview of our data collection framework, which comprises three main components. Firstly, we discuss the process of gathering educational and non-educational websites, as detailed in Sect. 3.1. Secondly, we present the methodology for scanning historical snapshots from Internet Archive's Wayback machine, which is described in Sect. 3.2. Finally, we discuss the approach for extracting third-party trackers from HTML files, which is outlined in Sect. 3.3.

3.1 Collecting Websites

To understand the evolution of web tracking in educational websites, we compare them to a control set of non-educational websites to see whether there are any changes related to education in particular. The comparison set is explicitly

controlled for popularity so that the two sets consisting of educational and non-educational websites have equal rank distribution. The studied websites must also have available historical data stored in internet archives.

We construct the two rank-matched sets of educational and non-educational websites as follows:

Step 1. We extract the educational websites from DMOZ[3]. DMOZ is a large communally maintained open directory that categorizes websites based on webpage content, and we use the DMOZ classification of educational websites. There are 146,941 websites in the DMOZ database labeled as educational websites.

Step 2. Next, we limit the set of educational websites to those occurring on the Open PageRank Initiative[4], which maintains a list of the top 10 million websites ranked based on their Open PageRank. There are 55,390 educational websites present among the top 10 million. This filtering is done so that we can create a comparable control set.

Step 3. We use the Internet Archive's Wayback Machine[5] for archived data. Therefore, the set of educational websites is further limited to those with at least one snapshot per year in every year from 2012 through 2021 to ensure that annual comparisons are balanced. This results in 17,975 educational websites altogether.

Step 4. Based on the list of educational websites from Step 3, we construct a set of non-educational websites with rank (Open Pagerank) distribution matching educational websites. Starting with the educational website of the highest rank, this is done recursively by choosing for each educational website a non-educational website that satisfies the following three conditions:

(a) The website has the lowest possible rank below the matching educational website.
(b) The historical data of the website is available on Internet Archive's Wayback Machine.
(c) The website has not already been added to the control set of non-educational websites.

For instance, if there are two educational websites of ranks 19 and 20, then ranks 21 and 22 would be chosen to the control set of non-educational sites, provided that they are not educational websites and have archived versions available. We study the rank gap (The rank of non-educational websites minus the rank of matching educational websites) distribution. The mean rank gap is 2.86, while the maximum gap is 255. We also find that 99% of the rank gap is below 16.

[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OMV93V.
[4] https://www.domcop.com/top-10-million-websites.
[5] https://archive.org.

3.2 Scanning the Historical Snapshot

There are two primary methods for scanning historical snapshots: the Wayback CDX Server[6] and the waybackpy Python library[7]. The Wayback CDX Server is a standalone HTTP servlet that serves the index used by the Wayback Machine to search for captures. The second method involves using the WaybackMachineCDXServerAPI provided by the waybackpy library to retrieve historical snapshots at specific times. For our research, we opted to use the Wayback CDX Server as our scanning method.

3.3 Extraction of Third-Party Trackers

Each website is examined for requests to other URLs initiated during the website's loading. These requests will always be embedded in three HTML-elements: *"iframe"*, *"script"* and *"img"*. We only consider the requests generated automatically without user action. That is why we omitted the *"a"*-element[8].

The list of third-party services (TPSs) was compiled by extracting all URLs found in the three HTML elements mentioned earlier across the entire dataset. For each website and URL, we checked whether the main domain of the linked URL (e.g., 'google' in 'www.google.dk') differed from the main domain of the website. If the domains were different, the URL was considered a 'third party' and the domain (e.g., 'google') along with the suffix (e.g., 'dk') were added to the list of TPSs.

To clarify our terminology, we will use the term 'trackers' instead of 'third-party services' for the remainder of this paper. While many third-party services serve various functions, such as providing weather data or chat services, some are solely designed for tracking and provide data that is used for personalized banner ads. However, even third-party services that seemingly provide non-tracking functionality have the potential to gather valuable data from users, such as their timestamped IP addresses and the websites they visit when the third-party service is activated. This information may be used by the third-party provider or sold to data brokers, or both. As all third-party services can track users, we refer to all such services invoked through websites as trackers [18].

We utilized the trackers list[9], which covers the period between May 2017 and August 2022 [13]. The trackers on the list are ranked according to their *tracker reach*, which is a metric defined in the aforementioned paper [13]. It should be noted that each tracker corresponds to multiple tracker domains. For

[6] https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server.
[7] https://akamhy.github.io/waybackpy/.
[8] A "ping"-attribute in the "a"-element allows requests to be made to multiple URLs without the user being aware of this, but there were no ping-attributes used in the data used in this study.
[9] https://whotracks.me/trackers.html.

instance, Doubleclick is associated with three tracker domains: '2mdn.net', 'doubleclick.net', and 'invitemedia.com'. In total, the tracker list comprises 1,285 tracker domains.

4 Analysis and Discussion

Our analysis focuses on the changes in web tracking between 2012 and 2021, with a particular emphasis on the qualitative and quantitative differences in tracker usage between educational and non-educational websites.

4.1 Evolution of Tracker Domains per Website

To begin our analysis, we computed the average number of trackers per website for each year. Figure 1 displays the trends in tracker usage on educational and non-educational websites between 2012 and 2021. In general, a striking 94.5% increase in the average number of trackers on educational websites was observed, while the control group experienced a comparatively modest 31.3% increase from 2012–2021. Specifically, when observing the trends of growth each year, we notice a plateau or slight reversal in growth occurring after 2017. Notably, the vertical line in Fig. 1 represents the formal implementation of GDPR in 2018. It is interesting to observe that the number of trackers on non-educational sites experienced a slight decline, whereas tracker usage on educational sites appeared to taper off around the same time.

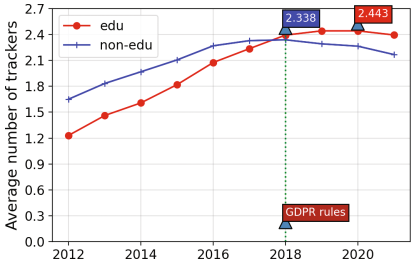
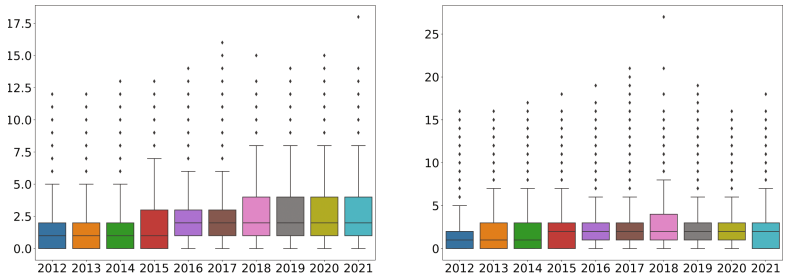


Fig. 1. Evolution of the average number of tracker domains per webiste.

Figure 2a shows a box plot of the number of trackers per year for educational websites. The plot suggests that the evolution in the average number of trackers after 2018 observed in Fig. 1 is driven by an increased dispersion in tracking across distinct educational sites, as the third quartile increases from 2017–2018, while the median (the horizontal line in each box) remains almost the same in the period 2017–2021. As a comparison, Fig. 2b shows a boxplot for non-educational websites. The plot also suggests the evolution in the average number of trackers in Fig. 1. The third quartile increased from 2017–2018 but has dropped since 2019. Especially, the first quartile decrease in 2021.



(a) educational websites. (b) non-educational websites.

Fig. 2. Number of trackers in each year for educational websites and non-educational websites.

According to the findings, it appears that users who browse educational websites are at a higher risk of having their online behavior information collected and potentially utilized by various services and websites. This disparity between educational and non-educational websites highlights the potential inadequacy of the GDPR in addressing privacy concerns specific to educational sites.

4.2 The Number of Trackers on Educational and Non-educational Websites

The tracking trends presented in Sect. 4.1 are purely descriptive, hence we conducted a statistical test to determine if there is a significant difference between educational and non-educational websites. Specifically, we performed a matched-pairs Wilcoxson signed rank(WS)[10] test to evaluate if the medians of the educational and non-educational samples are different for each year. This test is appropriate for paired data, which is the case for our study due to the rank-based construction of the data, and does not make any assumptions about the underlying distributions, making it a non-parametric test.

Table 1. Summary of the WS-test showing differences in each year, N = 17975.

Year	WS-test	
	p-value	Z
2012	5.2×10^{-86}	–19.66
2013	6.8×10^{-55}	–15.6
2014	1.6×10^{-46}	–14.32
2015	6.7×10^{-28}	–10.95
2016	7.1×10^{-11}	–6.52
2017	6.6×10^{-2}	–1.84
2018	1.5×10^{-5}	–4.33
2019	8.7×10^{-18}	–8.59
2020	3.2×10^{-20}	–9.21
2021	2.3×10^{-30}	–11.45

The input of the WS test is the number of tracker domains for each educational and non-educational website. The results of the tests are summarized in Table 1; as usual, small p-values indicate statistical significance; for all years, except 2017 $p < 0.01$

[10] http://www.biostathandbook.com/wilcoxonsignedrank.html.

for both tests. The sole exception is 2017, where $p > 0.05$ for the WS-test, consistent with the prevalence curves (see Fig. 1) crossing that year.[11]

4.3 Evolution of Usage Rate for the Most Common Trackers

To understand how tracking development differs between educational and non-educational sites, we compare how the ten most commonly occurring trackers have changed during the measurement period. We compute the usage of trackers based on the usage rate and select the top ten most used trackers in educational websites in 2012. The top ten trackers are on the vertical axis in Fig. 3.

We define the *usage rate* of each tracker as $f(t) = \frac{\mathrm{N}(t)}{\mathrm{N}(w)}$ where $\mathrm{N}(t)$ is the total number of websites where tracker t occurs, and $N(w)$ is the total number of websites in the sample. We calculate the relative increase I in usage rates of the top ten trackers most common on educational websites from 2012 to 2012 as $I = \frac{f(t)_{2021} - f(t)_{2012}}{f(t)_{2012}}$. The relative change of usage rate is shown in Fig. 3. We observe that the overall usage rate increased for the five top trackers on educational websites, including the social media sites Twitter and Facebook, and Youtube. It also increased for two Google-related trackers.

It decreased for five trackers, including Twimg (operated by Twitter) decrease by 66.4%, Addthis (–46.9%), Google-analytics (–32.5%), Adobe (–31.6%) and Googlesyndication (–1.1%) on educational websites. For non-educational websites, the usage rate increased only for the three Alphabet-operated trackers (Youtube, Googleapis and Google) and saw the largest decrease for Twimg, by 78.4%.

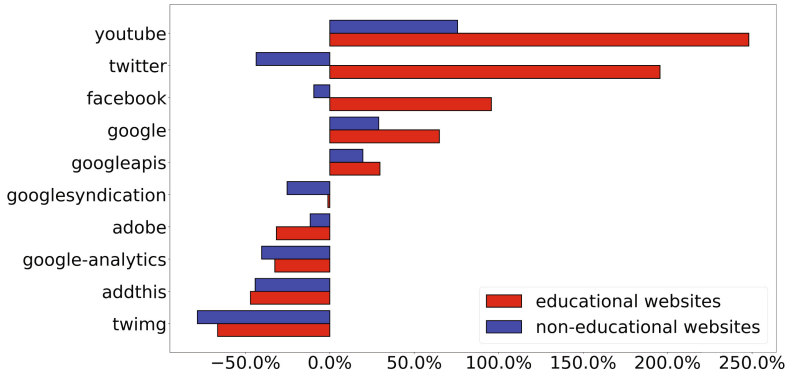


Fig. 3. Top usage rate change of trackers. X-axis is the percentage change of usage rate. Y-axis is the tracker name.

[11] Note that the level of statistical significance in the test means that correcting the alpha level for multiple comparisons does not alter the finding. This also holds for Table 2.

Notably, the use of Twitter, Youtube, and Facebook has evolved very differently between educational and non-educational websites. All three have an increased presence on educational sites, whereas their use has declined (Twitter and Facebook) or grown much slower (Youtube) on non-educational sites. The presence of trackers from these companies on educational sites helps target ads at the users when they visit the platforms and can also help educational sites to advertise their services to new, potential users with profiles similar to their existing users.

4.4 Distribution of Trackers in Educational and Non-educational Sites

The results in Sect. 4.1 show that the level of tracking differs significantly between educational and non-educational sites. We will investigate if the difference also relates to the composition of trackers used on the two groups of sites and differences in intensity. As stated in 3.3, all third-party services may collect data that can be used for tracking. However, the value proposition to the website owner differs between different services since they provide various functionalities to the site. Therefore, an analysis of the other functionalities also indicates what the site owner has sought to gain from embedding the service (or tracker), irrespective of the tracking of user behavior it enables. This analysis will, in turn, show if educational sites have followed a different path in integrating trackers than other sites.

For each year, we employ the two-sample Kolmogorov-Smirnov(KS) test–the standard non-parametric test for comparing distributions–to test whether the educational, resp. non-educational samples are drawn from the same underlying distribution. The test is suitable for paired data, similar to the WS test reported before. As indicated in Table 2, there is a significant difference in the distribution of trackers found on the two groups of sites in each year between 2012 and 2021. This indicates that in addition to the different quantitative trends, there appears to be a qualitative difference in the kind of trackers used on educational and non-educational websites.

Table 2. Summary of the KS-test showing differences in each year.

Year	KS-test	
	p-value	Statistic
2012	4.7×10^{-6}	0.25
2013	6.0×10^{-5}	0.22
2014	4.1×10^{-5}	0.23
2015	2.7×10^{-3}	0.18
2016	1.4×10^{-3}	0.19
2017	1.4×10^{-3}	0.19
2018	2.8×10^{-4}	0.21
2019	1.2×10^{-3}	0.19
2020	4.4×10^{-4}	0.2
2021	3.3×10^{-3}	0.17

4.5 Evolution of Different Categories of Trackers

To understand how the overall differences in tracker distribution identified in the previous section relate more closely to different purposes of web functionality, we look at the changes across different types of trackers. Since no exhaustive categorization of trackers exists, we use the tracker typology made available by

the WhoTracksMe initiative in June 2022, which to our knowledge, is the most comprehensive and up-to-date list, that is openly available[12].

This typology matches 1285 trackers in our dataset. While this represents only a subset of the total number of trackers, the list coincides with 213 of the 1285 most common trackers in our analysis. In the following, we examine the distribution of trackers across categories but only do so for the subset of the most common trackers. The WhoTracksMe list categorizes most common trackers into one of *Site Analytics*, *Customer Interaction*, *Advertising*, *Cdn*, *Social Media*, *Audio Video Player*, *Essential*, *Misc*. A more detailed explanation of the eight tracker categories is found in Appendix A.

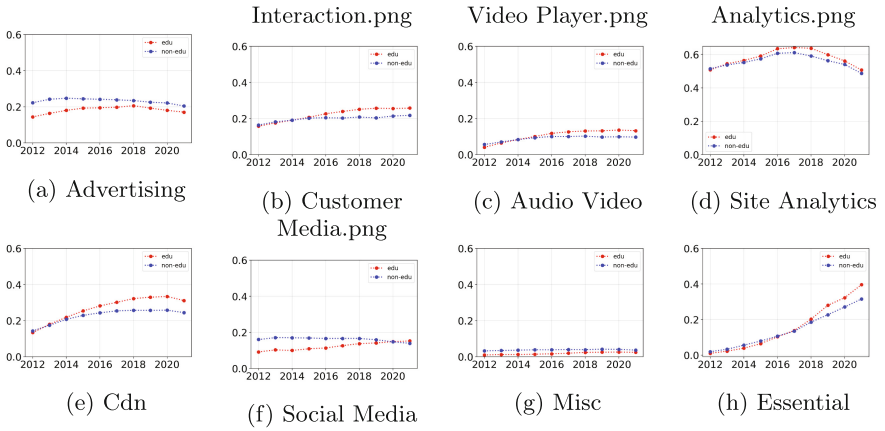


Fig. 4. Evolution of different categories of trackers from 2012 to 2021.

For each category c of trackers, we calculate usage rate $f(c)$ as $f(c) = \frac{\mathrm{N}(c)}{\mathrm{N}(w)}$ where $\mathrm{N}(c)$ is the number of websites this type of trackers occur on, and $\mathrm{N}(w)$ is the total number of websites in the sample. We calculated the usage rate $f(c)$ for each category c of trackers as $f(c) = \frac{\mathrm{N}(c)}{\mathrm{N}(w)}$, where $\mathrm{N}(c)$ is the number of websites on which this type of tracker occurs, and $\mathrm{N}(w)$ is the total number of websites in the sample. The evolution of different tracker categories in Fig. 4 indicates that, when comparing the levels in 2012 and 2021, educational sites have increased their use of all trackers except for *Site Analytics*. Even though *Advertising* and *Cdn* trackers have dropped slightly since their peak, they are still at a higher level than in 2012. In contrast, the usage rate for *Site Analytics*, *Advertising*, and *Social Media* is lower in 2021 than in 2012 for non-educational sites.

When comparing the two groups, two significant trends are apparent. Firstly, educational sites exhibit higher growth in the use rate of trackers related to interactive site features and audio-visual content. For instance, the *Audio Video Player* category witnessed a growth of 225.0% on educational websites from

[12] https://whotracks.me/trackers.html.

2012–2021 compared to 73.1% for non-educational websites. While the increase slowed down in 2018 or was even reverted for non-educational websites, it has increased again since 2019 and decreased since 2020. Additionally, *Cdn* services and *Customer Interaction* trackers have become more commonly used and grown more on educational sites than non-educational sites during the period. The increase from 2012–2021 was 63.7% for educational websites and 32.5% for non-educational websites in the case of *Customer Interaction* trackers. This growth in interactive features and audio-visual content on educational sites is consistent with the evolution of online learning, which has become more interactive and audio-visually engaging over the years [8]. Moreover, these trends make the sites more bandwidth-consuming, which is also in line with the growth of *Cdn* services.

Second, the use of *Social Media* (increased by 66.2%) and *Advertising* (18.0%) related trackers grow for educational sites but display a net drop for non-educational sites. Compared to the level in 2012, the use of both trackers is higher in 2021, whereas both are lower on non-educational sites. For both categories, the educational sites begin at a lower level than the non-educational sites. In both categories, the difference becomes less pronounced over time, and for *Social Media* ends up at the same level. This indicates that purely commercial tracking on educational sites has evolved from being comparatively less common than on other types of sites to be similar. For both types of sites, the use of trackers for *Advertising* has gone down in recent years, but for educational sites, the peak is more recent (2018) than for non-educational sites (2014). For non-educational sites, this is consistent with the overall development in commercial tracking, which has seen a general trend toward concentration around a few major players. In 2012, the market for commercial tracking was less dominated by monopolies such as Alphabet and Meta than it has since become [4]. The market domination of fewer players is consistent with the falling trend in the use rate. For educational sites, the continued growth is consistent with them catching up to the market standard for commercial tracking, which is also suggested by the strong growth of trackers from the top market players observed in Sect. 4.3.

4.6 Discussion

The evolution of web tracking over time aligns with predictions surrounding the implementation of GDPR in the EU, although its impact on educational sites has been less significant than on non-educational sites. While non-educational sites have experienced a decline in the use of trackers since the introduction of GDPR, the usage of trackers on educational sites has increased and remained stable at a higher level. Moreover, the prevalence of purely commercial tracking, such as advertising and social media tracking, has grown on educational sites, approaching similar levels as on non-educational sites.

The tracking via third-party services outside the commercial categories also serves other purposes (e.g., making it possible to embed a chat function on a web page). Any third-party service with a substantial use rate across the web gives the third-party company that operates the service the opportunity to collect

information about the end user. The overall increase in tracking also means that users of educational sites have become more exposed to having information about their online behavior collected and (potentially) exploited across a range of different services and sites. This development is interesting from a normative perspective: tracking on educational sites comes with specific privacy concerns since using these sites is not necessarily voluntary, and user consent to tracking is, therefore, less meaningful. Use of these types of sites happens both at different levels of the educational systems (schools and universities), and in the private sector, for example, as training of employees. The increased use of tracking, both through the inclusion of purely commercial trackers and by embedding other third-party services, suggests that learning activities are increasingly open for commercially oriented analytical exploitation.

The trends of tracking we have identified also suggest that the associated business model(s) remain active and are of increasing relevance in the online education sector. Our paper particularly raises the concern whether the GDPR in its current form suitably addresses privacy issues related to websites whose use is not predominantly voluntary, such as educational websites, but also many other sites like public websites, where the trends of tracking form an important research question on its right and should be addressed in studies in the future.

Reviewing the results, we do not find convincing evidence that tracking on educational websites has been substantially impacted by the COVID-19 pandemic. Despite the fact that additional traffic to these types of sites during the lock-down periods would represent a valuable asset for site owners, no trends in the data meaningfully relate to this. This may be related to the fact that educational sites had already adjusted their portfolio of commercial tracking in particular to facilitate monetization of increased traffic, e.g. through re-targeting of potential students on social media.

5 Conclusions

In this paper, we present a framework for examining historical web tracking within a defined set of sites, and apply it to a sample of educational websites. We constructed a sample comprising educational websites and a control group of non-educational websites that shared similar ranking positions. Utilizing the Internet Archive's Wayback Machine, we gathered historical data on third-party trackers. Our analysis involved 17,975 pairs of websites and their corresponding controls, spanning the period from 2012 to 2021. We observed a notable overall rise in tracking activities on educational sites.

Then we conduct a quantitative and qualitative analysis of third-party tracking on educational websites. We discover that the growth rate of educational websites has surpassed the control group from 2012–2021. Our investigation into the relative expansion of various tracker types suggests that the accelerated growth of tracking on educational websites may be attributed to the rising use of customer interaction, audio-visual content, and social media integration within these platforms.

Our analysis raises concerns about privacy and independence in education. Privacy issues in educational websites should be prioritized, as they can lead to unauthorized disclosure of confidential information, loss of trust, legal consequences, and intellectual property compromise. Furthermore, researchers may wish to analyze privacy lost in other areas, such as news or sports, from a historical perspective. Our framework offers a convenient solution for creating comparable websites and collecting historical third-party tracker data in these domains.

Appendix

A Tracker Categories

Trackers differ both in the technologies they use, and the purpose they serve. Based on the service they provide to the site owner, we have categorized the trackers in the following:
Advertising. Provides advertising or advertising-related services such as data collection, behavioral analysis or re-targeting.
Customer Interaction. Includes chat, email messaging, customer support, and other interaction tools
Essential. Includes tag managers, privacy notices, and technologies that are critical to the functionality of a website
Site Analytics. Collects and analyzes data related to site usage and performance. Social Media Integrates features related to social media sites
Audio Video Player. Enables websites to publish, distribute, and optimize video and audio content
CDN (Content Delivery Network). Content delivery network that delivers resources for different site utilities and usually for many different customers.
Misc (Miscellaneous). This tracker does not fit in other categories.
Essential. Includes tag managers, privacy notices, and technologies that are critical to the functionality of a website

References

1. Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., 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, pp. 674–689 (2014)
2. Agarwal, V., Sastry, N.: Way back then: a data-driven view of 25+ years of web evolution. In: Proceedings of the ACM Web Conference 2022, pp. 3471–3479 (2022)
3. Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., Mayer, J.: Privacy policies over time: Curation and analysis of a million-document dataset. In: Proceedings of the Web Conference 2021, pp. 2165–2176 (2021)
4. Bilić, P., Prug, T.: The Political Economy of Digital Monopolies: Contradictions and Alternatives to Data Commodification. Policy Press, Bristol (2021)

5. Binns, R., Lyngs, U., Van Kleek, M., Zhao, J., Libert, T., Shadbolt, N.: Third party tracking in the mobile ecosystem. In: Proceedings of the 10th ACM Conference on Web Science, pp. 23–31 (2018)
6. Binns, R., Zhao, J., Kleek, M.V., Shadbolt, N.: Measuring third-party tracker power across web and mobile. ACM Trans. Internet Technol. (TOIT) **18**(4), 1–22 (2018)
7. Dabrowski, A., Merzdovnik, G., Ullrich, J., Sendera, G., Weippl, E.: Measuring cookies and web privacy in a post-GDPR world. In: Choffnes, D., Barcellos, M. (eds.) PAM 2019. LNCS, vol. 11419, pp. 258–270. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15986-3_17
8. Elisabeta, P.M., Alexandru, M.R.: Comparative analysis of e-learning platforms on the market. In: 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), pp. 1–4. IEEE (2018)
9. Englehardt, S., Narayanan, A.: Online tracking: A 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1388–1401 (2016)
10. Hu, X., Sastry, N.: Characterising third party cookie usage in the EU after GDPR. In: Proceedings of the 10th ACM Conference on Web Science, pp. 137–141 (2019)
11. Jarke, J., Breiter, A.: The datafication of education. Learn. Media Technol. **44**(1), 1–6 (2019)
12. Jordan, K.: Degrees of intrusion? a survey of cookies used by UK higher education institutional websites and their implications. A Survey of Cookies Used by UK Higher Education Institutional Websites and Their Implications (March 16, 2018) (2018)
13. Karaj, A., Macbeth, S., Berson, R., Pujol, J.M.: Whotracks. me: shedding light on the opaque world of online tracking. arXiv preprint arXiv:1804.08959 (2018)
14. Kretschmer, M., Pennekamp, J., Wehrle, K.: Cookie banners and privacy policies: Measuring the impact of the GDPR on the web. ACM Trans. Web (TWEB) **15**(4), 1–42 (2021)
15. Krishnamurthy, B., Wills, C.: Privacy diffusion on the web: a longitudinal perspective. In: Proceedings of the 18th International Conference on World Wide Web, pp. 541–550 (2009)
16. Krupp, B., Hadden, J., Matthews, M.: An analysis of web tracking domains in mobile applications. In: Hooper, C., Weber, M., Weller, K., Hall, W., Contractor, N., Tang, J. (eds.) WebSci 2021: 13th ACM Web Science Conference 2021, Virtual Event, United Kingdom, 21–25 June 2021, pp. 291–298. ACM (2021)
17. Lerner, A., Simpson, A.K., Kohno, T., 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)
18. Libert, T.: Exposing the hidden web: an analysis of third-party http requests on 1 million websites. arXiv preprint arXiv:1511.00619 (2015)
19. Mathur, A., et al.: Dark patterns at scale: Findings from a crawl of 11k shopping websites. In: Proceedings of the ACM on Human-Computer Interaction 3(CSCW), pp. 1–32 (2019)
20. Sanchez-Rola, I., et al.: Can i opt out yet? GDPR and the global illusion of cookie control. In: Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, pp. 340–351 (2019)
21. Saxena, A., Saxena, P., Reddy, H., Gera, R.: A survey on studying the social networks of students. arXiv preprint arXiv:1909.05079 (2019)

22. Sørensen, J., Kosta, S.: Before and after GDPR: the changes in third party presence at public and private European websites. In: The World Wide Web Conference, pp. 1590–1600 (2019)
23. Urban, T., Tatang, D., Degeling, M., Holz, T., Pohlmann, N.: Measuring the impact of the GDPR on data sharing in ad networks. In: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, pp. 222–235 (2020)
24. Vlajic, N., El Masri, M., Riva, G.M., Barry, M., Doran, D.: Online tracking of kids and teens by means of invisible images: COPPA vs. GDPR. In: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, pp. 96–103 (2018)