Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites

Jannick Sørensen
Aalborg University Copenhagen
Copenhagen, Denmark
js@cmi.aau.dk

Sokol Kosta
Aalborg University Copenhagen
Copenhagen, Denmark
sok@cmi.aau.dk

ABSTRACT

The commencement of EU's General Data Protection Regulation (GDPR) has led to massive compliance and consent activities on websites. But did the new regulation result in fewer third party server appearances? Based on an eight months longitudinal study from February to September 2018 of 1250 popular websites in Europe and US, we present a mapping of the subtle shifts in the third party topology before and after May 25, 2018. The 1250 websites cover 39 European countries from EU, EEA, and outside EU, belonging to categories that cover both public-oriented citizen services, as well as commercially-oriented sites. The developments in the numbers and types of third party vary for categories of websites and countries. Analyzing the number of third parties over time, even though we notice a decline in the number of third parties in websites belonging to certain categories, we are cautious about attributing these effects to the general assumption that GDPR would lead to less third party activity. We believe that it is quite difficult to draw conclusions on cause-effect relationships in such a complex environment with many impacting factors.

CCS CONCEPTS

• **Information systems** → **Web mining**; **Traffic analysis**; **Information retrieval**; *Data mining*; • **Security and privacy** → Social aspects of security and privacy.

KEYWORDS

GDPR, Web measurement, Third-Party web server, Tracker, Public Services, General Data Protection Regulation

ACM Reference Format:
Jannick Sørensen and Sokol Kosta. 2019. Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313524

1 INTRODUCTION

Mainly driven by privacy concerns, much research has been dedicated on examining the presence and implications of contacts with third-party servers when users browse websites. EU's General Data Protection Regulation (GDPR[1]) aims to give users more control over the collection and distribution of their personal information, also on websites. As one sign of the relevancy of GDPR to websites, the European advertiser's lobby organization (IAB Europe[2]) has proactively assisted and informed their members on the implications of GDPR [14], e.g. by providing an open source consent framework. Before the enforcement of GDPR, the future of third parties and advertising was also a hot debate topic in the advertising industry [5, 17, 26, 36]. As every provider serving users from the EU were compelled to review their use of cookies and other person-identifiable data-exchange with third parties, the assumption in the industry was that GDPR would lead website providers to review and reduce the number of third party servers [25]. Correspondingly, a number of companies started to offer the monitoring of third party activity on behalf of website providers, as well to legalize the data collection by obtaining user consents. The much ado in the web industry in relation to GDPR provokes thus the question whether the activity level of third party HTTP responses changed with GDPR. Do users meet fever third party servers when browsing the WWW?

This paper presents the results of an ongoing longitudinal study of strategically selected websites, from European countries in and outside the EU/EEA, and from USA. The websites have been selected to represent those categories that have been found to show either very high or very low numbers of third party[3] URLs [10]. Earlier works have observed that sites which belong to government organizations, universities, and non-profit entities have a low number of TPs, while sites with editorial content have a high number of TPs, since *"they are pressured to monetize page views with significantly more advertising"* [10]. In this work, we examine this observation with a longitudinal dataset spanning across the commencement of GDPR and across different countries.

Our intuition is that the picture is more differentiated than what has been studied so far, as many high-traffic public organizations' websites, such as public service media companies or public transportation companies, feature advertising and are deeply integrated in media ecosystems [22, 30, 37]. We analyze if these citizen-serving websites may thus, with respect to the number of TPs, have similarities with commercial high-traffic websites. As an example, public service media organizations promote themselves as *"islands of trust and quality in the multimedia environment"* [7]. They argue that citizens have high level of trust in public service media [8, 9]. Arguably, websites published by public entities should be expected to protect users' privacy particularly well, in order to maintain

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313524

[1] EU General Data Protection Regulation (GDPR) – https://eugdpr.org/
[2] IAB Europe – https://www.iabeurope.eu/
[3] In the rest of the paper we refer to third parties as "TPs".

the integrity of the public organization as being independent and sovereign in relation to private actors. As public organizations are funded by tax or – in cases of public service media – via a license fee, involvement with advertising blurs the public/private distinction. Indeed, advertising in public service broadcasting is subject to heavy regulations [33]. In a digital context, users may pay twice for the same public service: first via the tax and then with their data being transferred to the third party web services. The extended discussion of public organizations' use of third party services is however not the aim of this paper, but to test our assumption of websites morally obliged to a low number of third party contacts, we include law firms as examples of private websites that could have an interest in signalling a high level of privacy.

Our contributions: Having collected HTTP requests and responses 21 times over eight months for a strategic selection of 1250 European and US websites, we can analyze the fluctuations in number of TPs before and after the commencement of GDPR on May 25, 2018. Thereby, we provide empirical evidence for the discussion whether GDPR would lead to fewer TPs. At a general level, we can conclude that the amount of TPs on web pages have slightly declined, but the picture is more complex and contradictory when we study the developments for respectively categories of sites and TPs. This finding adds details to findings provided by [21] and [10]. We cannot support the general assumption that the GDPR has led to fewer TPs, since we cannot find strong evidence for any correlation. As a second contribution, we characterize the differences between websites offered by public and private organizations from the TP perspective, initiating a discussion on public organizations' use of TPs. Finally, we intend to publish the data to allow other research groups further analysis.

2 RELATED WORKS

The involvement of third-party (TP) advertisers or trackers in websites and mobile apps have interested many researchers [1, 2, 6, 10, 18, 20, 29, 31, 32, 34, 35]. In 2014, Acar et al. presented one of the first large-scale studies of third-party trackers in web pages, showing that *the web never forgets* [1]. The authors show that trackers use different and complex ways to integrate in a website, so that it is difficult to distinguish them accurately from genuine content. In a more recent work, Lerner et al. perform an *archaeological analysis* of tracking practices in the web, analyzing websites from 1996 to 2016 from the Internet Archive[4], showing that the first web trackers date back to 1996 [18]. Unsurprisingly, the study finds that the number and complexity of TPs have increased since their first appearance in 1996. Furthermore, through a large-scale automated analysis of third-party privacy policies, Libert et al. show that reading and understanding privacy policies is far from easy, and users are not presented with the needed information allowing them to take relevant decisions on what data to share [20].

One of the largest measurements of third-party presence was conducted in January 2016, where Englehardt and Narayana developed OpenWPM[5], an automated script that uses Firefox web-browser to visit websites through the Selenium[6] automation tool. OpenWPM collects and stores several browsing data, such as HTTP requests and responses, in a database for later analysis. The authors used the tool to visit one million unique websites via 90 million web page requests [10]. The selected websites represented the one million most popular websites in the world, according to the internet measuring service Alexa[7]. Following the categorization of websites made by Alexa, the authors found that news websites are among those with the highest number of third-parties, while websites from public organizations, NGOs, and universities are among those websites with the lowest number of third-party web-server contacts. In this study, the authors found over 81,000 unique third-party servers. However, they notice that only 123 of these were present at more than 1 percent of the visited sites [10].

Many studies have used OpenWPM for crawling and analyzing the web. Among these, Binns et al. assess the power of third party trackers through business collaboration [2], Mazel et al. compare privacy protection methods [23], Brookman et al. analyse cross-device tracking [3], and recently, Rodriguez et al. have analyzed tracking cookies in the EU [27].

To understand the inter-dependencies in the business ecosystem of media production, Lindskow analyzed the TPs presence at 41 US-American news publishers, identifying 1,356 business partners [22]. The study shows that *digital news publishers function as integrators of resource flows from hundreds of business partners of different types and sizes and that each of these partnerships involves complex resource exchanges and strategic dependencies.* The author's focus is thus not on privacy, but on business networks. However, they provide useful insights into dividing the TPs in different categories, identifying and analyzing six categories, as follows: *i)* News content elements, *ii)* News content, *iii)* Editorial tools, *iv)* Measurement, *v)* Advertising, and *vi)* Supporting resources.

Starov et al. analyzed the identifiers that third-party trackers store in users' devices to uniquely distinguish between users, showing that there are many websites that use the same identifiers, allowing the authors to cluster websites that looked unrelated to each other [32]. Falahrastegar et al. [12] make an early contribution to map the anatomy of the third party ecosystem, with a focus on user tracking.

Recently, with the advent of GDPR, many researchers have started analyzing the possible effects that GDPR is having on the web. Libert et al. [21] present a fact-sheet on third-party cookies of 204 news websites from seven EU countries. The pages were visited in April and July 2018 (pre- and post-GDPR), using a crawling system developed by one of the authors in a previous work [19]. The authors observe that the percentage of web pages that contain third-party content decreases with only two percent from April to July 2018, dropping from 41 to 40 third-parties per page. Conversely, they notice that there is a drop of 22 percent in the number of cookies, with variations among the seven countries.

Very recently, Iordanu et al. performed a massive analysis of a large-scale traffic flow from final users to third-party trackers [16]. The authors implemented a browser extension and distributed it to 350 real users, creating this way a reliable dataset. Then, they generalized their analysis methodology by extending it to datasets from four large Internet Service Providers (ISPs) containing data

[4] https://archive.org/
[5] https://github.com/citp/OpenWPM
[6] https://www.seleniumhq.org/
[7] https://www.alexa.com/topsites

from more than 60 million users. The study was performed during a period of four months before the implementation of GDPR. Among other results, this work shows that around 3% of the flowing traffic is related to sensitive user data, which might violate the GDPR requirements. Nevertheless, the authors show that trackers collecting sensitive data of users in the EU keep them primarily within the EU (in 90% of the cases).

In this work, we perform a large-scale analysis of third-party HTTP responses within a period of time that covers the pre- and post-GDPR era, specifically from February to September 2018. Firstly, we manually select more than 1300 websites of different categories from 39 countries, so that we have a reliable and controlled coverage of countries of interest and website purpose. Then, we use OpenWPM [11] to automatically visit the websites and some randomly selected sub-pages, creating a dataset composed of 21 crawls/harvests[8]. We then analyze the possible effects of GDPR, by comparing the changes of the TPs characteristics in the visits performed before and after GDPR.

3 RESEARCH DESIGN

In this section, we give a description of the tools we use to perform the data collection and the methodology we follow for selecting the websites to analyze, the cleaning of the data, and the classification of the third-parties.

3.1 Selection and classification of websites

The search for websites to be analyzed reflects our strategic selection of those website categories where third-party trackers have shown either high or low interest.

3.1.1 Identifying website categories. Considering the analysis and results of Englehardt et al. [10] and Libert et al. [21], we look for types of websites where either a high or a low number of TPs should be expected. From these studies, we can observe that there is a significant correlation between the ownership of the website, as being either private or public, and the number of TPs [10]. Indeed, while a user of commercially-oriented websites may expect a high presence of third party web servers, a user of a publicly owned citizen-oriented websites should expect none or very few third parties. As such, our website selection process reflects the distinction between websites offered by **public** and **private** organizations as a first criterion.

Then, the second selection criteria is the **country of origin** of the websites. We target primarily websites from the European countries, distinguishing between the EU, EEA (Extended Economic Area[9]), and non-EU sites. However, we also include few websites located outside Europe, mainly from US. To identify the country of a website we use its country code top-level domain, and if that is not available we perform a WhoIs look-up[10] to find the registrant organization's country.

Then, we identify eleven **categories of websites** that people visit more often, starting from and extending the list proposed in [10]. So, we include examples of shopping sites, hotel, travel and holiday sites, gaming and gambling sites, video-on-demand sites,

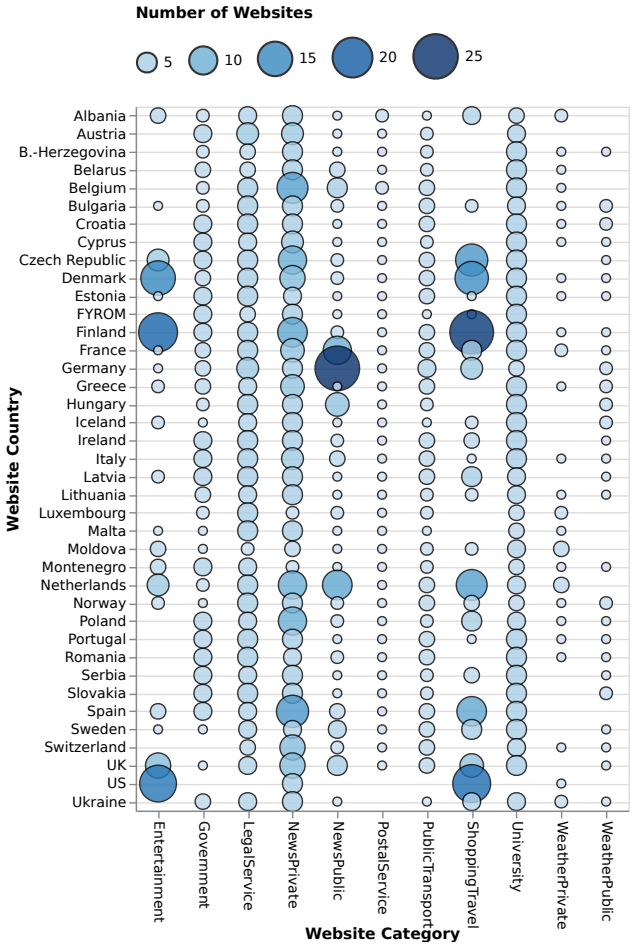

Figure 1: Number of websites per country and category selected for the analysis in this work.

general entertainment, and life-style magazines. In the analysis we have collected all these sub-categories in a single main category "Shopping and Travel". Moreover, we collect websites from European countries from law firms and attorneys, assuming that we would find an example of private websites with a low number of TPs. Furthermore, we consider governmental organizations, universities, public transportation, postal services, and state-funded weather services as public services websites, assuming that we would not observe a high number of TPs. For governmental organizations we specifically selected ministries and public administration sites.

Finally, we consider the news websites, which, differently from previous works, we explicitly divide them into public and private. We make this distinction considering the fact that there are huge differences among news/media organizations, particularly in Europe, where some are publicly owned and financed via a license fee, tax and in some cases also advertisement, whereas others are privately owned and financed by subscription, advertisement and in some cases public support. Similarly, we also distinguish between public weather websites from governmental/public organizations,

[8] In this paper, we use the terms *crawl* and *harvest* interchangeably.
[9] www.efta.int/eea
[10] https://whois.icann.org/en

and websites from private providers. To the best of our knowledge, this is the first work that makes this clear distinction.

As such, we identify the following categories: *i)* Entertainment, *ii)* Government, *iii)* Legal Services, *iv)* News Private, *v)* News Public, *vi)* Postal Services, *vii)* Public Transport, *viii)* Shopping and Travel, *ix)* University, *x)* Weather Private, and *xi)* Weather Public.

3.1.2 Selection of websites. To find the relevant websites for each category, we have used various methods: via search engines, via user involvement, via related research, and via lists of membership. Differently from previous works, which use Alexa[11] to classify a website as belonging to a category, we manually look at each website and classify it with human intervention. We believe that, compared to Alexa's lists, the manual classification allows us to have a higher level of certainty about the websites' category, given that we do not have access to Alexa's criteria for the categorization.

To identify public service media news websites we use the list of members of the European Broadcasting Union, a lobby organization for public service media[12]. As for private media, we take the top-5 most popular news sites listed in the Reuters Institute Digital News Report 2017 [24]. For countries not covered in the Reuter's report, we use the Alexa web service[13]. For websites in other categories than news, we manually identify them using Google Advanced Search[14], filtering by region/country, getting the top 5 websites not already in our list. Moreover, to broaden the collection of websites further, we perform a web survey asking users from our personal network to mention the most popular websites from their countries in the following categories: news, lifestyle/magazine, shopping, travel & hotels, entertainment, and gaming. From 17 users we obtained 279 unique websites: 62 news sites, 42 magazine/lifestyle sites, 67 shopping sites, 62 travel & hotel and accommodation sites, 23 entertainment sites, 20 gaming sites, and 3 sites outside these categories.

Overall, we were able to carefully select a total of 1,363 websites from 39 different countries, divided into 11 categories, as shown in Figure 1. As can be observed from the distribution, we are aware that not all categories are equally well represented for all countries, as well as some categories are represented with more websites in some countries than in others. This bias is an implication of our main study goal, which aims at having at least five websites from each country in these categories: *i)* private news, *ii)* universities, *iii)* ministries, *iv)* law firms. As for the categories of public transportation, postal service, and public weather, for some countries we found fewer sites than in some other. For public service media we included all public service organizations, even if multiple as in the case of Germany. The variability in the other categories, i.e. Entertainment and Shopping & Travel, is then due to user involvement.

3.1.3 Selection of subpages. To ensure a broad capture of possible TPs on the websites we have generated a list of subpages to visit. We have done so by collecting random URLs from the 1,363 sites, producing in total +50,000 URLs. From these, we randomly select ten subpages to be visited for each site. On average, we have visited 9.35 pages from each site. Some sites do not have subpages, since one single page provides all the necessary information. This is the case e.g. for weather information and public transport (journey planners and timetables). In total, the script visited 12,778 subpages from the 1,363 websites for each of the 21 harvests, capturing all the HTTP requests and responses. For each website and subpage visit we collect the set of unique third party URLs.

[11]https://www.alexa.com/topsites
[12]https://www.ebu.ch/about/members
[13]https://www.alexa.com
[14]https://www.google.dk/advanced_search

3.2 Data collection and cleaning

3.2.1 Sampling period. We selected a long sampling period for the large-scale crawling, in order to be able to detect the effects of GDPR in terms of possible and expected deviations in the number of different types of TPs on the different categories of websites over time. As such, the sampling period spans across the GDPR commencement, which was on May 25, 2018, with 8 harvests from February to May and 13 harvests from May 25 to September.

3.2.2 Website harvesting. To perform the automatic website crawling we setup a virtual machine (VM) with 5 CPUs and 8GB of memory running Ubuntu. We installed the widely used OpenWPM framework and configured it to use four vanilla Mozilla Firefox browsers in parallel, meaning that we did not in any case register or log in with a user account while visiting the websites. Moreover, we did not give consent to any storage of cookies or other types of privacy-related requests, as the interaction with the web pages only took place automatically via the script. In this way, the behavior of the script consequently imitates a user who ignores all user privacy consent messages from the websites. However, we did not install any extension on the browser that would stop the websites from storing or reading cookies. We did not erase any cookie after a harvest was performed, so that on next visits the websites could behave normally by accessing the stored information. Finally, it is worth mentioning that the physical machine on which the VM was deployed is located in the EU, meaning that the visited websites must comply with GDPR.

3.2.3 Data cleaning and selection for analysis. Through the 21 harvests, we collected in total 31,493,575 HTTP responses from the 12,778 subpages of the 1,363 websites, averaging 1,431,526 HTTP responses per harvest.

Before analyzing the data, we performed a cleaning process that led to 83 websites to be removed from the dataset. The reasons for excluding these URLs are the following: *i)* Some of them redirected to a website already in our collection, thus risking to produce duplicated entries; *ii)* Some of them, suggested by the survey users, were from countries outside Europe (Japan, Korea, Hong Kong, Panama), from which we have no other or very few websites; *iii)* Finally, to consolidate our eleven main analytic categories of websites (see Figure 1), we have discarded few websites that did not fall into any of the main categories.

Thus, in the rest of the paper we analyze and present our findings regarding 1,250 websites from 39 countries from the EU, EEA, countries outside EU/EEA, and from the USA, and divided into 11 categories, as shown in Figure 1.

3.3 Categorization of the TPs

In this section we present some results related to the overall number of third-parties observed in our datasets and to the methodology we

TP Category	No. of Unique URLs
Advertising	1382
Publisher	341
Distribution Technology	237
Analytics	72
Retail	170
Content	209
Programming	82
Plug-in	45
Social media	30
Editorial	31
Privacy	6
Search Engine	27
Cybersecurity	16
Unidentifiable	347

Table 1: Number of unique TP URLs for each TP category.

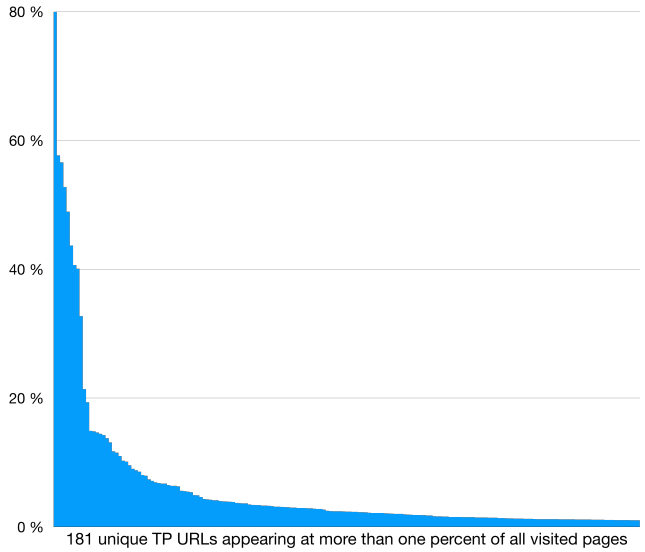

Figure 2: The presence in percent for the most omnipresent TPs of all pages.

pursued on classifying them into different categories. To identify the third-party URLs we look at all HTTP responses and extract all those URLs that are different from the website URL that was being visited. From this analysis we find 3,128 unique third-party URLs.

Then, we manually categorize the TPs by visiting the unique URLs. If we find a readable website, we attempt to determine the type of services offered by the TP by reading their descriptions of services, products, and mission statements. If the original TP URL redirects to another website, e.g. a company website, we analyze the description of products and services of that company. Otherwise, if the URL is not accessible and returns an error message, such as *HTTP 404 Not Found* for example, we classify the TP URL as *Unidentifiable*. Following this procedure, we were able to identify with high accuracy the 14 TP categories listed in Table 1, where we also show the number of TPs per each category. The details of each category are the following:

Advertising contains all the various services related to sale of ads and analysis of user profiles, such as: Programmatic advertising services, Personalization services, Recommender systems, Real-time bidding platforms, Demand-side and Sell-side platforms, Data management platforms and data integration, Data brokers and data trading, Re-targeting systems, User trackers, Ad-servers, Advertising agencies, Ad-verification systems, Brand-integrity, attribution and anti-fraud systems, Marketing automation, Content & native advertisement services, Cross-device user identification systems, and Video-based advertising.

Analytics contains different types of services used to understand user behavior and gather user feedback: Audience measurement, AI-powered analysis of user behavior, Audience Intelligence, Semantic profiling, Audience research (qualitative), Customer flow, Marketing analytics, Quality of Service monitoring and Web performance optimization, Attention optimization tools for Publishers, and Customer feedback.

Content includes all types of elements shown on the web page, not being advertising. That includes content embedded from other websites, not part of the media company/organization.

Cybersecurity specifically focuses on services that perform internet infrastructure surveillance.

Distribution technology includes content delivery networks, cloud services, and streaming services.

Editorial contains services specifically aimed at editors, e.g. recommender systems designed for publishers.

Plug-in contains web services that integrate content from other services into the visited site.

Privacy contains services that monitor website compliance with GDPR and cookie use on the visited site.

Programming contains services that deliver code to the rendering of the web page.

Publisher encompasses all servers that are owned by the media organizations including collaborating media organizations.

Retail includes all-purpose web-portals, job-seeking portals, shopping platforms, real-estate brokers, and consumer products (advertisers).

Unidentifiable includes those URLs that do not return a valid and readable HTML page when visited in the browser, but a *404* or *forbidden* error message or a blank page. Our working assumption is that a large portion of the unidentifiable TPs are advertising related, delivering the ads to be displayed, measuring and verifying the exposure, or tracking the user across websites and other devices. This assumption is based also on an expert interview that we conducted with the CTO of an advertising technology company in December 2017[15].

4 RESULTS

Our research design allows analysis of the collected data along different dimensions. Indeed, the categorization of both the visited websites and third party URLs, combined with 21 samples (harvests)

[15] Jacob Knobel, former CTO of Densou.dk, also cited in [30]

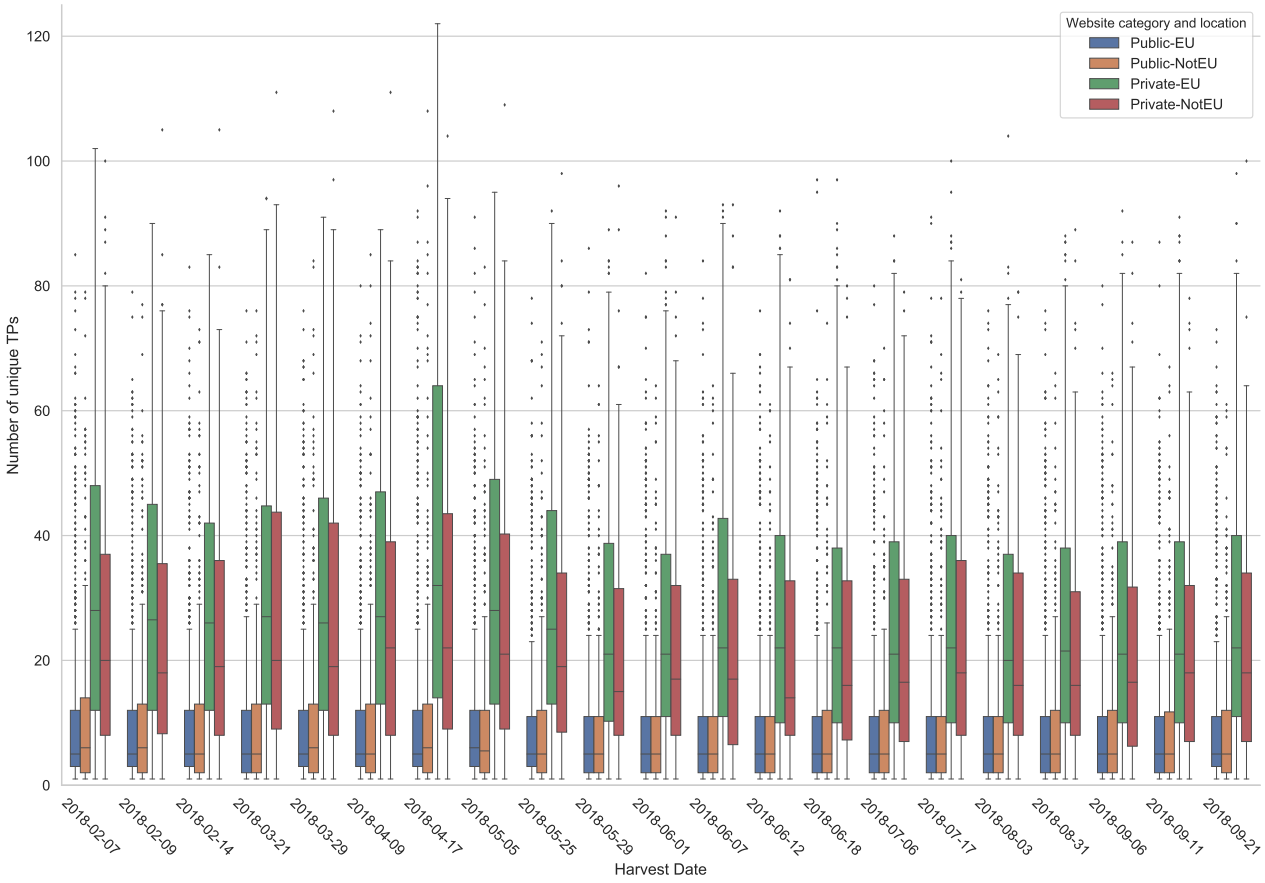

Figure 3: Number of TPs with respect to the website being public or private and its EU membership (EU/NotEU). In this graph we consider EEA countries together with the EU ones.

over a time span of eight months, gives us the opportunity for a detailed investigation over different feature combinations.

Firstly, as the sampling period spans across the commencement of GDPR, we can compare pre- and post GDPR levels for the presence of TPs. Moreover, the categorization of the visited sites into countries, groups of countries, and types of site, together with the categorization of the TPs, allows us to create a detailed picture of the developments. Furthermore, we can look at groups of sites, such as EU and EEA countries compared with sites from outside the EU/EEA area, and extend the investigation by also differentiating the websites in private and public ones. Finally, on a more detailed level, we can also look at the variations in the number of TPs for different categories of sites, for different countries, and for different types of TPs.

4.1 Analyzing the Number of TPs for Categories of Sites

In total, we found 3,128 unique third party URLs, appearing on at least one web page in all our 21 harvests. During all the observations, we found that:

- After May 25, 675 unique TPs disappeared, of these 63 were present in all harvests before May 25.
- 644 new unique TPs appeared after May 25, of these 36 were present in all harvests after May 25.
- 2096 TPs were present in all 21 harvest on at least one website.
- 408 TPs appeared only in one harvest at one website.

If we consider a list with the top-20 TPs, we find similarities with earlier works showing the dominance of the giant companies [10, 21, 28]. From our data, we find that this includes nine TP URLs controlled by Google *(google-analytics.com, doubleclick.net, googleapis.com, google.com, gstatic.com, google.dk, googletagmanager.com, googlesyndication.com, googletagservices.com, googleadservices.com)*, two TPs controlled by Facebook *(facebook.com, facebook.net)*, Amazon's CDN *(cloudfront.net)*, and the competitor CDN *(cloudflare.com)*, the advertising companies Adnexus *(adnxs.com)*, *criteo.com*, *adform.net*, the analytics companies *scorecardresearch.com (TMRG) and gemius.pl*, and the omni-present *twitter.com*. These URLs represent the beginning of a very long tail where only 151 TPs have a share of one percent or more of the pages, while the remaining 968

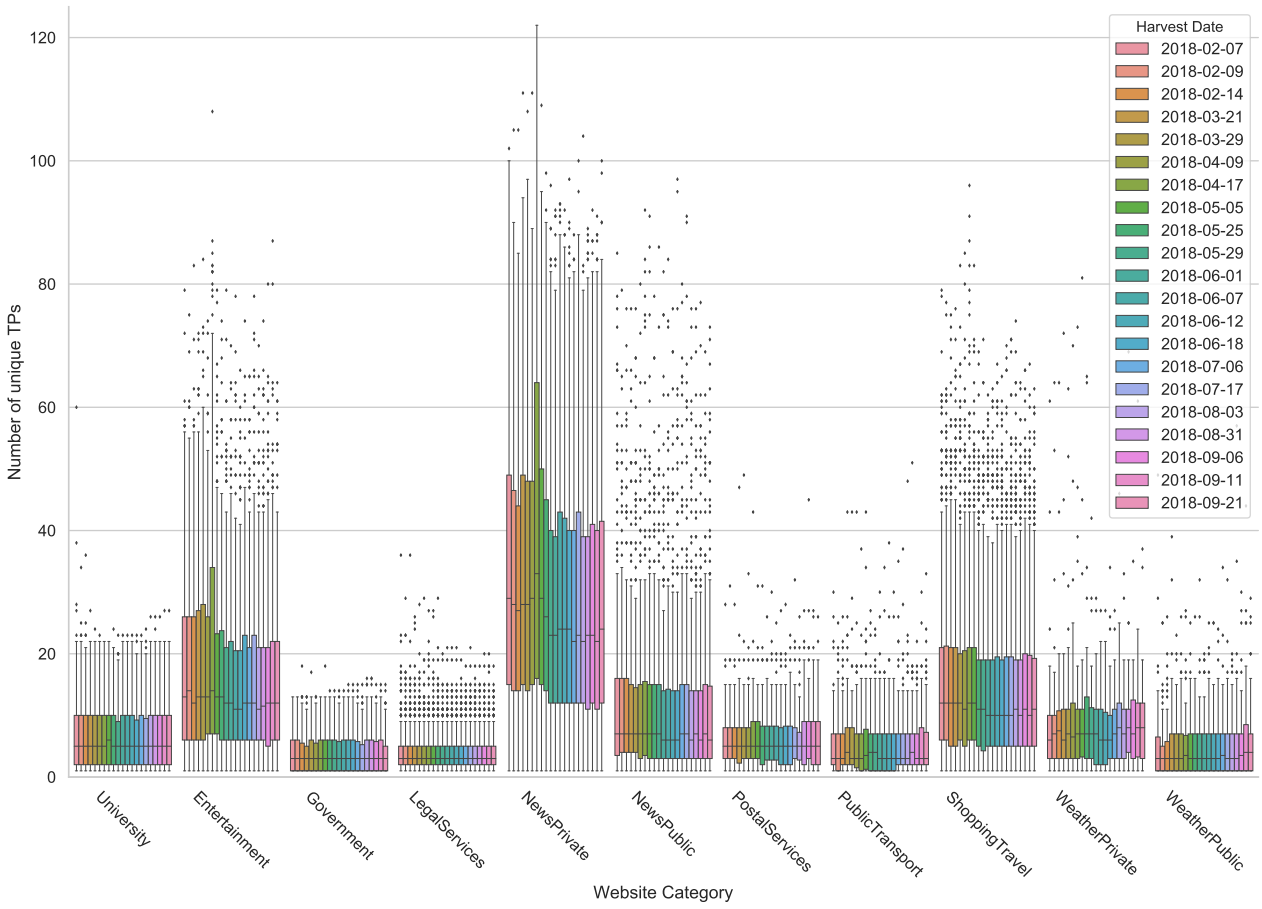

Figure 4: Number of TPs per website category with respect to the harvest date.

TPs have a share of less than one percent. The cut long tail of the top 151 TPs is presented in Figure 2.

In Figure 3, we analyze the variations of the TPs by dividing the websites into private/public and EU/NotEU, where EU in this case includes also the EEA countries, while *NotEU* includes websites from countries outside the EU/EEA. The figure shows, for each harvest, the distribution of the number of TPs embedded in each website represented as box plot, showing the minimum, the median, the maximum, the first and third quartile, and the outliers. From the figure, we can firstly see that the **public websites have far fewer TP URLs than the private ones** in all the harvests. Secondly, we can see that there is a **difference between private EU and NotEU websites**, while there is not much difference when considering the public websites. When considering the variations over the observed time, we can see a rather **stable picture for the public sites**, regardless being from the EU or not. In comparison, the **private sites present large fluctuations**, but with fewer outliers. Moreover, these data suggest that the **private websites present a slight decrease of the number TPs** after the commencement of GDPR, while the public ones are hardly affected.

Next, we continue our investigation trying to identify the responsible private websites that cause the decrease in the number of TPs after GDPR. In Figure 4, we present the distribution of the number of TPs for website category, grouped by harvesting date. As also observed above, this figure shows clearly and with more details that **public websites** (i. e. categories *University, Government, PublicTransport, PostalServices, NewsPublic,* and *WeatherPublic*) **present a low level of TPs and a low level of fluctuations**, with the exception of some outliers. Of these public categories, the news websites present the highest number of TPs and the highest variations. When it comes to **private websites**, *LegalServices* present a similar constant low level of TPs as the public ones, followed by the *WeatherPrivate*, while ***Entertainment, NewsPrivate,* and *ShoppingTravel* present the highest number and variations of TPs**.

To quantify the variations over time, we perform the linear regression of the average number of TPs for all website categories for each harvest. The regression lines are presented in Figure 5 and more details are presented in Table 2. From these results we can confirm that the ***NewsPrivate, Entertainment,* and *ShoppingTravel* categories show a clear decline in average numbers**

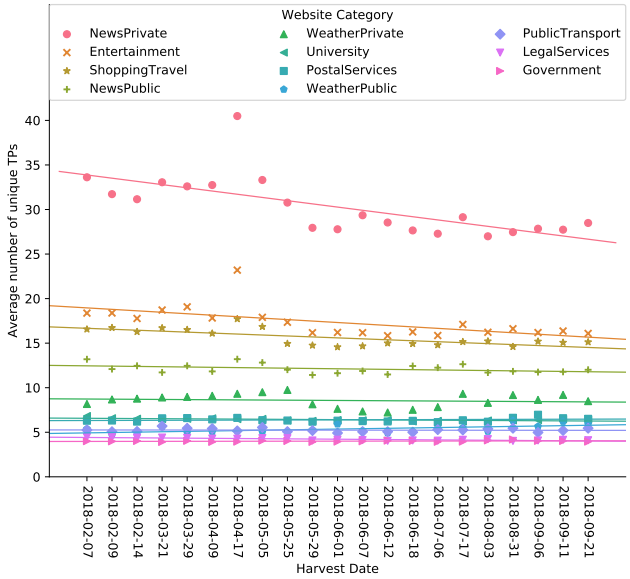

Figure 5: Average number of TPs per each website category during the harvesting period. The line represents the linear regression of the number of TPs, showing the tendency of the changes. The slope coefficients and the intercepts for each line are given in Table 2.

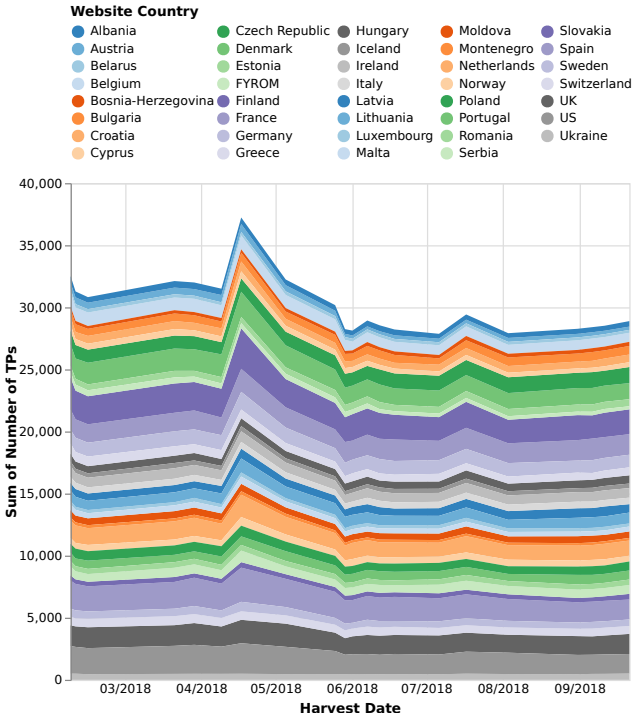

Figure 6: The total number of unique TPs at websites from 39 countries over time.

Private/Public	Website Category	Slope	Intercept
Private	NewsPrivate	-0.360	35.0
Private	Entertainment	-0.164	19.5
Private	ShoppingTravel	-0.107	17.0
Private	LegalServices	-0.019	4.5
Private	WeatherPrivate	-0.016	8.8
Public	NewsPublic	-0.033	12.6
Public	University	-0.015	6.6
Public	PublicTransport	-0.002	5.3
Public	Government	+0.003	4.0
Public	PostalServices	+0.008	6.3
Public	WeatherPublic	+0.043	4.8

Table 2: Coefficients of the regression lines of Figure 5.

of unique TPs per site visit, while for the other categories the decline is small. Moreover, for three categories of public websites, namely *PostalServices*, *Government*, and *WeatherPublic*, we see a slight, although we believe insignificant, growth in the number of unique TPs per site.

4.2 Countries Perspective

As a next analysis, we consider the individual countries and try to see the effects of GDPR over time. In Figure 6, we present the developments in the total number of unique TPs for each site we have visited from the different countries. This graph shows again that overall there has been a slight decline of the total number of TPs, but does not easily allow for a detailed quantification of the changes for each country. For this reason, we calculate for each country the average number of TPs for the harvests performed before and after GDPR and represent the changes as a percentage in Figure 7. The results show big variation among countries, with Estonia having the biggest growth and Austria having the biggest decline. 21 EU countries have fewer TPs after GDPR, while 7 EU countries have more TPs. Both EEA countries, Norway and Iceland, present a decline. Among the eight European countries outside EU/EEA, seven have a decline, while only Bosnia-Herzegovina has a growth.

4.3 Analyzing the distribution of the TPs

In this section, we present the analysis of the distribution of TPs in the different categories and the changes in each category over time (refer to Section 3.3 and Table 1 for more details about TP categorization). In Figure 8, we show a stream graph of the total number of TPs observed during each harvest. From the figure we identify major fluctuations in the *Unidentifiable* and *Advertising* third party categories. This is obviously due to the fact that these are the biggest categories, so it is easier to spot the changes.

To quantify the variations as we did for the countries, we look at the changes in percent and compare the visits after GDPR with those from before. The results for the categories with a declining trend are the following: *Cybersecurity (-51,16%), Privacy (-34,07%), Unidentifiable (-18,56%), Analytics (-11,28%), Advertising (-11,26%), Retail (-10,34%), Content (-6,50%), Distribution technology (-5,91%),*

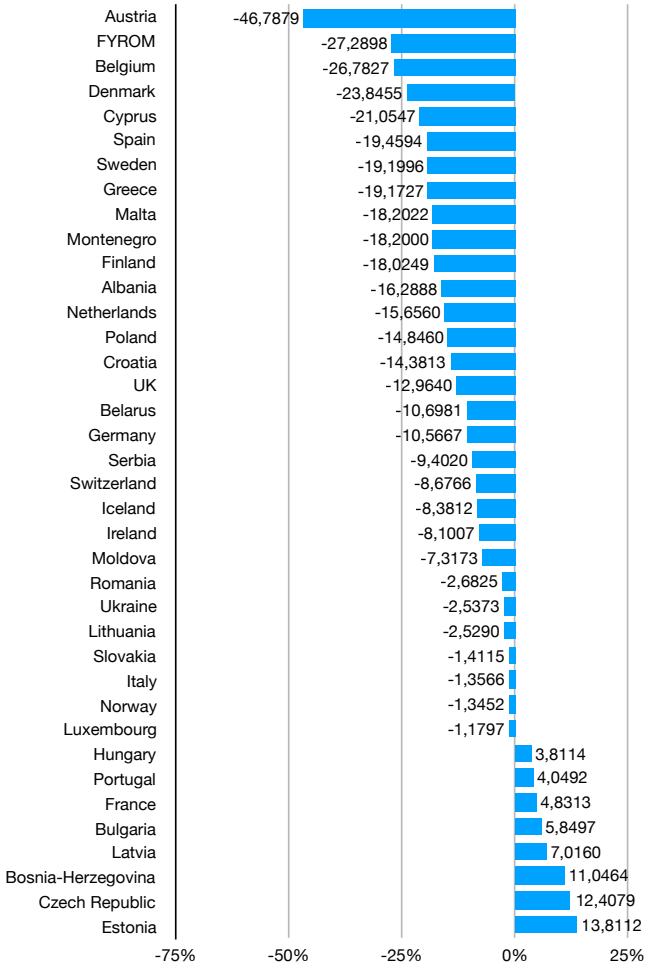

Figure 7: Comparing pre- and post May 25 average numbers of unique TPs for countries, listed as change in percent.

Social Media (-5,79%), *Publisher (-5,77%)*, and *Plug-In (-2,07%)*. We see a growth only in the categories *Programming (+2,73%)* and *Search Engines (+8,32%)*. However, as some of the categories have very few unique TPs the above changes should be considered with care. When considering single websites, we noticed that not all websites have a fluctuating number of TPs: for 241 websites (91 private and 148 public), the number of TPs is constant for all visits.

4.4 The Relation between Categories of Websites and TPs

The advantage of categorizing both websites and TPs is that we can also examine their mutual relation. Figure 9 shows, with a logarithmic representation, the occurrence of unique TPs of different categories in different categories of websites. The heat map considers all visits during all the harvests aggregated. The **most intense relationship can be seen in the lower right corner, where private media and advertising intersect**. We can also notice that the presence of advertising-related TPs is quite high

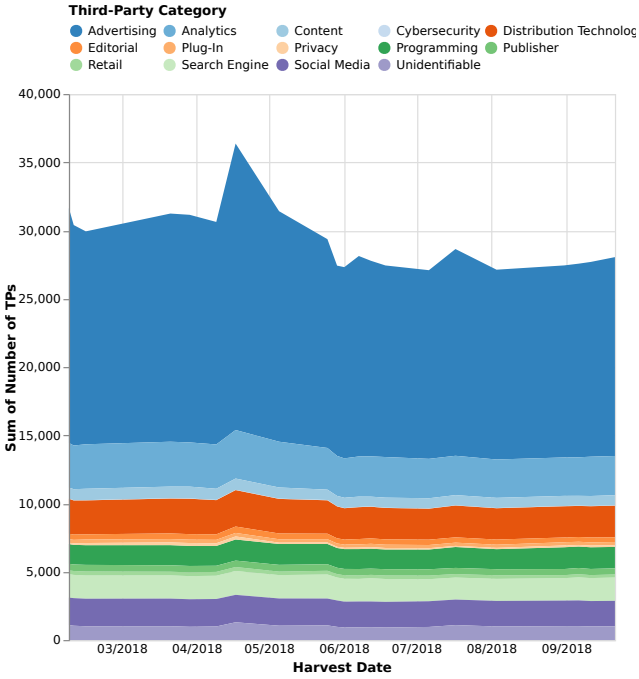

Figure 8: The total number of TPs distributed in categories.

in all the other types of sites, especially on public service media pages, on shopping/travelling, and entertainment sites. We can also notice that some TP categories are only found at some types of visited sites: The TP categories *Editorial* and *Publisher* are unique to the media websites. Other categories, such as *Cybersecurity* and *Privacy*, are low both in the number of unique TPs and in distribution and can thus not be expected to be prominently represented in the heat map. As a potentially more controversial finding, we see occurrences of both TP categories *Retail* and *Advertising* being present in the website categories *University*, *Government*, *WeatherPublic*, *PostalServices*, and *PublicTransport*. That indicates that on the level of third-party ecosystems, **the borders between public and private institutions are to a certain degree dissolved**, with a possible user-tracking across the public-private border.

4.5 Interpretation

The different categories of websites have both *i)* a different behavior during the sample period and *ii)* a different pattern in the distribution of websites with many or few TPs. When we again look at Figure 4, we can see that the number of TPs for some categories of websites remains unchanged over the sampling period, while others have fluctuating amounts of TPs. Further detailed analyses on the level of individual pages and individual TPs may clarify the pattern. By looking at Figure 4, we can also see that the categories are showing a different span between the website in the category with the fewest TPs, and the website in same category with the most TPs. For websites in categories such as *Entertainment, NewsPrivate, NewsPublic*, or *ShoppingTravel*, this calls for a differentiated analysis. Possibly, for each of the categories websites with respectively few

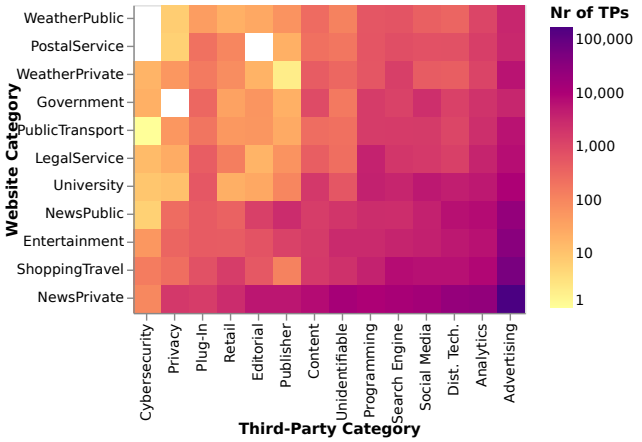

Figure 9: The intensity of the relation between categories of websites and categories of TPs on a logarithmic scale.

or many TPs may have more in common (also TPs) with websites in other categories. Again, this calls for an analysis of the appearance pattern for individual TPs across websites. An assumption, that can be supported by our findings, is that the fluctuations are mostly related to advertising and unidentifiable TPs, as shown in Figure 8.

Looking at the amount of TPs over time, we cannot see a decline for most categories of sites, and for others only a small decline. The assumption that GDPR would lead to a remarkable decline cannot be supported, as the fluctuations are too many for most of the categories of sites. Only for the categories *NewsPrivate*, *Entertainment*, and *ShoppingTravel* we could see a significant decline. However, to interpret this as an effect of GDPR may be problematic, since drawing strong conclusions on cause-effect relationships in such a complex environment with many impacting factors, such as online advertising and web page compilation, is difficult.

One hypothesis could be that advertisers prefer to reduce the amount of TPs to ease the obstacle of obtaining consent. Another hypothesis could be that the decline is caused by a technological change in the advertising industry. Indeed, the advertising industry is currently changing the methods for conducting the real time bidding for the sale of advertisements on web pages [15]. The old method conducts the bidding from the user's browser, producing - in the case of a difficult bidding process - a large number of HTTP requests and responses to real-time bidding platforms and their associated *Supply-side* and *Sell-side* platforms, among other third parties [4, 13]. The new server-side header bidding does not involve the user's browser, but takes place between the publisher website (the first party) and the servers of the bidding platforms. While the privacy exposure may be as big as before, the user will experience fewer HTTP requests from her browser [15]. Further research is however needed to assess the uptake of server-side header bidding by publishers and other websites displaying advertisements. Based on our current research it is thus problematic to associate the decline in advertising-related and unidentifiable TPs with the introduction of GDPR.

5 DISCUSSION

Our analysis shows that the number of third party servers a user will meet at different websites in many cases fluctuates from visit to visit. It is thus difficult to draw the conclusion that GDPR should have resulted in a user meeting fewer third parties when browsing a website. Furthermore, attributing the general decline we see to GDPR is problematic, as the implementation of GDPR coincides with another change in the ecosystem of third party services, namely the uptake of server-side header bidding for online advertisements [15] (see the previous section). Further analyses is however needed to shed light on these changes.

One can interpret the third party activities in different ways. The literature has mainly looked at these from a privacy perspective, focusing particularly on the tracking of users across websites. Through our categorization of third parties, based on text analysis of services offered by the third party company, we can however see that user tracking, segmentation, and user profiling have become standard products offered for sale to any advertiser. That does not make the privacy issue less pressing, but it shows that tracking has become industrialized.

Another way to interpret the third party activities is to see them as sign of an ever more integrated world wide web; a business ecology of services that co-produce the experiences to users, an interpretation aligned also with previous works [22]. That approach raises two discussions: one on *dependence* and one on *efficiency*.

The dependence discussion, initiated by Lindskow et al. [22], focuses on media companies' dependency on external collaboration in order to reach their audiences and monetize the contact. The outsourcing of production of the website content through a deep integration in the third party service ecology, diminishes the power and freedom of the media institutions. They operate on top of an infrastructure composed by a complex network of third party services, a network which has become the gateway to users.

The efficiency is a rather technical discussion on improving page loading times, initiated by the advertising industry. The answer has been to conduct the bidding for advertisements on the server-side, between the media organization's server and the bidding platforms with their affiliated network of services. The study of third-party interaction in the browser may however in the future not provide us much information of the exchange and analysis of our user data.

6 CONCLUSIONS AND FUTURE WORK

Our longitudinal large-scale study of the third-party server interactions at websites has shown that no clear effect of GDPR can be seen. Fluctuations of the number of TPs during the eight months harvesting period can be observed for most categories of websites, with some categories, such as private news, shopping/travel, and entertainment, presenting a visible decline. Other categories have a stable or even slightly growing number of unique third-party web services. The detailed study on the level of different countries adds nuances to the distinction between EU/EEA countries and other non-EU countries and the USA. Particularly, we see a difference between private websites from within the EU/EEA and from outside. Furthermore, our research shows large differences between privately owned websites and publicly owned websites, with the

former having many TPs embedded and the latter having low but growing amount of TPs.

As future work, we intend to analyze the relations between single third party servers and single websites, possibly employing machine learning techniques to uncover hidden patterns and relations. Furthermore, we will also continue harvesting the selected websites, potentially also extending the list. Finally, we will also focus on further analyzing the utilization of third party services in publicly owned websites.

REFERENCES

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The Web Never Forgets. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, Gail-Joon Ahn, Moti Yung, and Ninghui Li (Eds.). ACM Press, New York, New York, USA, 674–689. https://doi.org/10.1145/2660267.2660347

[2] Reuben Binns, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2018. Measuring third party tracker power across web and mobile. (2018), 1–21. https://doi.org/10.1145/3176246 arXiv:1802.02507

[3] Justin Brookman, Phoebe Rouge, Aaron Alva, and Christina Yeung. 2017. Cross-Device Tracking: Measurement and Disclosures. *Proceedings on Privacy Enhancing Technologies* 2017, 2 (2017), 133–148. https://doi.org/10.1515/popets-2017-0020

[4] Oliver Busch. 2016. The Programmatic Advertising Principle. In *Programmatic Advertising: The Successful Transformation to Automated, Data-Driven Marketing in Real-Time*, Oliver Busch (Ed.). Springer, Cham, 3–15. https://doi.org/10.1007/978-3-319-25023-6_1

[5] Yuyu Chen. 2018. GDPR is coming, and data management platforms are in the crosshairs. https://digiday.com/marketing/gdpr-coming-data-management-platforms-crosshairs/

[6] Mark D. Corner, Brian N. Levine, Omar Ismail, and Angela Upreti. 2017. Advertising-based Measurement: A Platform of 7 Billion Mobile Devices. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (MobiCom '17)*. ACM, New York, NY, USA, 435–447. https://doi.org/10.1145/3117811.3117844

[7] EBU Digital Strategy Group. 2002. *Media with a purpose - Public Service Broadcasting in the digital era. The Report of the Digital Strategy Group of the European Broadcasting Union.* Technical Report Version DSG 1.0. European Broadcasting Union, Geneva. http://www.ebu.ch/CMSimages/en/DSG_final_report_E_tcm6-5090.pdf

[8] EBU Media Intelligence Service. 2017. *Market Insights : Trust in Media.* Technical Report. European Broadcasting Union, Geneva.

[9] EBU Media Intelligence Service. 2018. *Market Insights: Trust in media 2018.* Technical Report. European Broadcasting Union, Geneva. 40 pages. https://www.ebu.ch/files/live/sites/ebu/files/Publications/MIS/login_only/market_insights/EBU-MIS-TrustinMedia2018.pdf

[10] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. ACM, New York, NY, USA, 1388–1401. https://doi.org/10.1145/2976749.2978313

[11] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Extended Version of the paper presented at ACM CCS 2016.* http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

[12] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. 2014. Anatomy of the Third-Party Web Tracking Ecosystem. (sep 2014). arXiv:1409.1066 http://arxiv.org/abs/1409.1066

[13] Oliver Gertz and Deirdre McGlashan. 2016. Consumer-Centric Programmatic Advertising. In *Programmatic Advertising The Successful Transformation to Automated, Data-Driven Marketing in Real-Time*. 55–73. https://doi.org/10.1007/978-3-319-25023-6_5

[14] IAB - Interactive advertising bureau Europe. 2018. GDPR Implementation. https://www.iabeurope.eu/category/policy/gdpr-implementation/

[15] IAB - Interactive advertising bureau Europe. 2018. *Header Bidding and Auction Dynamics.* Technical Report August. interactive advertising bureau Europe. https://www.iab.it/wp-content/uploads/2018/09/IAB-Europe_Header-Bidding-and-Auction-Dynamics-White-Paper_August-2018-1-compressed.pdf

[16] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing Cross Border Web Tracking. In *Proceedings of ACM IMC 2018.* Boston, MA.

[17] Robin Kurzer. 2017. What does the GDPR mean to your third-party data processors? https://martechtoday.com/gdpr-mean-third-party-data-processors-208098

[18] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. 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)*. USENIX Association, Austin, TX. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/lerner

[19] Tim Libert. [n. d.]. webXray. https://webxray.org/

[20] Timothy Libert. 2018. An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 207–216. https://doi.org/10.1145/3178876.3186087

[21] Timothy Libert, Lucas Graves, and Rasmus Kleis Nielsen. 2018. *Changes in Third-Party Content on European News Websites aft er GDPR.* Technical Report August. Reuters Institute, Oxford University. 1–7 pages. https://reutersinstitute.politics.ox.ac.uk/our-research/changes-third-party-content-european-news-websites-after-gdpr

[22] Kasper Lindskow. 2016. *Exploring Digital News Publishing Business Models - A Production Network Approach.* Ph.D. Dissertation. Copenhagen Business School. http://hdl.handle.net/10398/9284

[23] Johan Mazel, Richard Garnier, and Kensuke Fukuda. 2017. A comparison of web privacy protection techniques. (2017). https://doi.org/10.1016/j.devcel.2015.10.024.Role arXiv:1712.06850

[24] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy, and Rasmus Kleis Nielsen. 2017. *Reuters Institute Digital News Report 2017.* Technical Report. Reuters Institute for the Study of Journalism, Oxford. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/DigitalNewsReport2017web_0.pdf

[25] Louise Reseke. 2018. JP/Politiken tager opgør med annonce-jungle og dropper 200 samarbejder. https://mediawatch.dk/secure/Medienyt/Web/article10631483.ece

[26] Patricio Robles. 2018. GDPR: What future for first, second and third-party data. https://econsultancy.com/gdpr-what-future-for-first-second-and-third-party-data/

[27] Elsa Rebeca Turcios Rodriguez. 2018. *Tracking Cookies in the European Union, an Empirical Analysis of the Current Situation.* Master. Delft University of Technology.

[28] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12)*. USENIX Association, Berkeley, California, 12–21.

[29] Franziska Roesner, Chris Rovillos, Alisha Saxena, and Tadayoshi Kohno. 2015. TrackingObserver: A Browser-Based Web Tracking Detection Platform. https://trackingobserver.cs.washington.edu

[30] Jannick Kirk Sørensen and Hilde Van den Bulck. 2018. Public service media online, advertising and the third-party user data business. *Convergence: The International Journal of Research into New Media Technologies* (aug 2018). https://doi.org/10.1177/1354856518790203

[31] Bharat Srinivasan, Athanasios Kountouras, Najmeh Miramirkhani, Monjur Alam, Nick Nikiforakis, Manos Antonakakis, and Mustaque Ahamad. 2018. Exposing Search and Advertisement Abuse Tactics and Infrastructure of Technical Support Scammers. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 319–328. https://doi.org/10.1145/3178876.3186098

[32] Oleksii Starov, Yuchen Zhou, Xiao Zhang, Najmeh Miramirkhani, and Nick Nikiforakis. 2018. Betrayed by Your Dashboard: Discovering Malicious Campaigns via Web Analytics. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 227–236. https://doi.org/10.1145/3178876.3186089

[33] Torben Stühmeier and Tobias Wenzel. 2012. Regulating Advertising in the Presence of Public Service Broadcasting. *Review of Network Economics* 11, 2 (jan 2012). https://doi.org/10.1515/1446-9022.1251

[34] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J. Weitzner, and Nigel Shadbolt. 2017. Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5208–5220. https://doi.org/10.1145/3025453.3025556

[35] Tim Wambach and Katharina Bräunlich. 2017. The Evolution of Third-Party Web Tracking. In *Information Systems Security and Privacy. ICISSP 2016. Communications in Computer and Information Science*, Olivier Camp, Steven Furnell, and Paolo Mori (Eds.), Vol. 691. Springer, 130–147. https://doi.org/10.1007/978-3-319-54433-5_8

[36] Chris Ward. 2018. Will GDPR kill the third-party data market? https://www.mycustomer.com/marketing/data/will-gdpr-kill-the-third-party-data-market

[37] Michele Zanitti, Sokol Kosta, and Jannick Sørensen. 2018. A User-Centric Diversity by Design Recommender System for the Movie Application Domain. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1381–1389. https://doi.org/10.1145/3184558.3191580