Privacy Policies Caught between the Legal and the Ethical

European Media and Third Party Trackers before and after GDPR

Sørensen, Jannick Kirk; van den Bulck, Hilde; Kosta, Sokol

DOI (link to publication from Publisher):
10.2139/ssrn.3427207

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Sørensen, J. K., van den Bulck, H., & Kosta, S. (2019). *Privacy Policies Caught between the Legal and the Ethical: European Media and Third Party Trackers before and after GDPR*. Paper presented at TPRC47: Research Conference on Communications, Information and Internet Policy, Washington DC, Washington, United States. https://doi.org/10.2139/ssrn.3427207

Privacy Policies Caught between the Legal and the Ethical: European Media and Third Party Trackers before and after GDPR.

Jannick Kirk Sørensen (Ph.D.)
Center for Communication, Media and Information Technologies (CMI)
Aalborg University of Copenhagen
js@cmi.aau.dk

Hilde Van den Bulck (Ph.D.) (corresponding author: hdv26@drexel.edu)
Dep of Communication Studies
Drexel University

Sokol Kosta (Ph.D.)
Center for Communication, Media and Information Technologies (CMI)
Aalborg University of Copenhagen
sok@cmi.aau.dk

DRAFT PAPER: DO NOT QUOTE WITHOUT PRIOR CONSENT FROM THE AUTHORS

Privacy Policies Caught between the Legal and the Ethical: European Media and Third Party Trackers before and after GDPR.

INTRODUCTION

This contribution analyses the use of third-party services by media websites in 39 European countries before and after the introduction of the EU's General Data Protection Regulation (hereafter: GDPR) May 25, 2018 as an inroad to discuss the legal versus ethical obligations of legacy broadcast media with regards to audiences' privacy and the impact on the key value of trust in media.

The media ecosystem appears in a never-ending state of flux, following ongoing technological, economic and socio-cultural developments. Especially the move to web-based media services, including social media, have pushed legacy media and law and policy makers to keep up. Recent years have seen them struggling with issues revolving around technological innovations that provide media with new opportunities for content creation and dissemination as well as for audience relations but that come with their own set of legal and ethical issues (Van den Bulck et al., 2018a). This is especially relevant for public service media (PSM), that are expected to maintain particular standards and values in return for their continued existence and that do not easily fit the new media reality. For one, diversifying audience behaviour requires new ways to track their movements and, thus, to reach and retain them. Efforts to do so are part and parcel of the so-called datafication of the media-ecology whereby all online human action (clicks, likes, survey results) is tracked and translated into quantified and quantifiable 'big data' (Mayer-Schoenberger & Cukier, 2013, Van Dijck, 2014). To help their internet-based content reach audiences, media collaborate with outside parties: when a user loads a webpage, a number of external services are contacted, triggered by scripts embedded in the received webpage or in other scripts. Such 'third-party' services range from helping editors, marketing people, media researchers and PSM managers to analyse user behaviour; providing technical help, for instance with the distribution of heavy data like video streams; showing content from social media or advertising.

These developments have created shifts in legacy media's value chains and business models (Raats et al., 2015; Donders & Van den Bulck, 1016; Donders et al., 2018.) and raise important ethical and legal questions as, potentially, these third parties can identify users and collect user data that are of use to these media but are also tradeable, i.e. of interest to other parties. This creates potential threats to citizens' rights leading to discussions about privacy and surveillance (Srnicek, 2017; Zuboff, 2019). This is a concern for all media as their success and survival relies on maintaining respectful relations with their audiences. However, it is of particular relevance to PSM, the raison d'être of which is based in their role and position as trusted institutions (Van den Bulck & Moe, 2018). Law and policy makers, from their end, have been trying to deal with these issues of privacy and surveillance as, increasingly, critical voices speak up, demanding action. The most forcible reaction to date has come from the European Union who has stepped in with its General Data Protection Regulation (GDPR). In force since 25 May 2018, GDPR deals with the third-party servers (represented by URLs) that play a role in compiling a webpage presented to the user and aims to provide extended rights to users to protect personal information by giving users more control over the collection and distribution of their personal information, also on websites. GDPR forces every website provider serving users from the EU 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 (Reseke, 2018), while legal experts suggest that the principles of purpose limitation as well as retention minimization inherent in GDPR may, in fact, enable so-called Big Data (Mayer-Schönberger & Pandova, 2016). This results in our broad research question: *Do (legacy) media audiences meet fewer third party servers when accessing their digital offerings than before GDPR*?

Beyond this legal question, though, the issue involves an ethical principle. As audiences, citizens, civil society and policy makers voice concerns, much of the public and political outcry has been focusing on large social media and digital giants like Facebook and Alphabet, with eye catching cases such as Cambridge Analytics paving the way. However, notions of ethics extend to broadcast and other (legacy) media, especially PSM. For instance, where print and, especially, free to air broadcast media mainly had anonymous listeners and viewers, online media can - per definition - not avoid knowing their audience, at least at the very basic level of the users' IP-addresses contacting the server of the media website. The technical structure of web communication invites for a much more detailed study of users, dissolving the original anonymity associated with broadcasting. This creates two problem areas for legacy broadcast media, especially PSM.

One potential problem area relates to media respecting their audience's privacy, as tracking and data mining can occur without users' knowledge, infringing their privacy. This is of particular relevance to PSM as their brand identity and reputation (and public financing), to a large extent, are based in a role and position as 'trusted institutions'. Different from the audience anonymity in broadcasting, for interactive digital platforms, the burden of demonstrating integrity and independence falls back on the PSM organisations, in terms of ensuring the anonymity of users and their browsing behaviour in relation to external 'third-party' web services.

The other area of concern legacy broadcast media's relationship to online advertising. Advertisers have come to expect state-of-the-art segmentation of users to ensure the highest possible impact of the advertising messages, making so-called 'programmatic advertising' (Busch, 2018) the dominant type of interactive advertising. However, this requires advertising technology such as user-tracking across platforms; user interest profiling; browsing history analysis for advertising retargeting and 'intention to buy' calculations for advertising price setting. As a result, web browsing has evolved into a complex interaction between, often, hundreds of servers each time a user visits a page. The question is whether legacy broadcast media are vigilant in their use of third party services and whether GDPR has worked as a wake-up call for all legacy and especially public service media. This is of particular relevance for PSM as it involves their relationship to commercial revenue, in most countries a debated and regulated issue.

To this end, after this introduction, this contribution first develops in a theoretical framework that combines an understanding of general principles of and existing research into third party server use with conceptual insights into the relationship between the legal and the ethical, based in two complementary theoretical perspectives: media values, especially PSM values, and computer ethics. Next, turning to the empirical section, we discusses the methodological set-up of our data collection and analysis. Subsequently, the results section discusses occurrences and distribution of TP URLs across various categorisations, i.e. according to different media and different types of third party servers.Cluster analysis reveals four clusters of media according to number and types of third parties involved.Crucially, the result sections compares these data before and after the introduction of GDPR. Finally, the conclusion reflects on these results in light of the theoretical framework, weighing legal versus ethical consideration in how media deal with third party servers, and zooming in on the implications for media, especially PSM.

THEORETICAL FRAMEWORK

Third-party servers on media content websites

Attention grabbing cases such as the Snowden revelations and Cambridge Analytics as well as more critical voices warning for hypes such as FaceApp that creates an image of your older self in return for your data, have brought to the general attention the widespread use and pitfalls of third-party servers in Big Data, now part and parcel of our online behaviour (Lyon, 2014; Dolata, 2017). Third-party servers (represented by URLs) play different roles in compiling the webpage presented to the user: some deliver videos, images or sound streaming, others give technical resources like computer code to build webpages. We focus on third-party servers that track, collect and analyse user behaviour to report site traffic to editors or to optimize exposure to content and advertising. Advertising-related third-party servers are a booming industry, delivering advertising banners, video and native advertising; segment and profile users by tracking them across platforms; auction personalized web advertisement slots to advertisers; and 'retarget', i.e. show the user a 'reminder advertisement' from a previously visited website. To refine their services, they subject large amounts of user data to advanced, artificial intelligence-based analysis to predict future behaviour like buying intention or to describe interests in narrow user segments. The more precise the description of a user, the higher the potential value of an advertising slot/inventory (Turow, 2011). The monetary value hereof is significant. In 2018, online advertising for 27 countries in Europe (incl. Russia and Turkey), was a €55.1 billion business - compared to €34.0 billion on TV ads and €23.3 billion on print media ads (IAB Europe, 2019) – and $107.5 in the US, up from 88 billion in 2017 (IAB, 218a). Typically, data management platforms offer user data for a price to other actors. Information regarding the value of user profile information is hard to obtain, but in 2009 the price of a targeted ad was 2.68 times the price of an untargeted ad (Beales, 2010, cited by Acquisti, Taylor and Wagman, 2016: 24).

Research into third-party servers

The activity of third-party servers has caught researchers' interest. Some authors (Falahrastegar et al., 2014; Wambach and Bräunlich, 2017) focus on the technologies that identify users in the browser from one website visit to the next, either through 'cookies' (Internet Engineering Task Force, 2011) or through so-called 'finger-printing' technologies that identify users across devices without cookies (Acar et al., 2014). Other studies map and categorise which third-party sites are contacted when users visit webpages. For instance, using automatic scripts, Englehardt and Narayanan (2016) analysed the one million most popular websites in the world in 2016. Visiting the pages 90 million times in one month, they found 81.000 different third-party URLs, yet only 123 were present at more than one per cent of the websites. Their study further showed that news websites have the highest average number of third-party URLs, while government, non-profit and university organizations' websites have the lowest number. A study (Sørensen & Van den Bulck, 2018) comparing unique URLs found on PSM and commercial broadcast media's web pages during 6 visits between December 2016 and August 2017, i.e. before GDPR, showed an average of 42,95 third-party URLs among private media compared to 70,42 for PSM with advertisements, 37,60 for PSM with possibility for advertisements, and 17,33 for PSM not allowed to display commercial advertisements. While low numbers of third-party URLs for PSM could be related to a ban on commercial advertisements, private media, too, showed considerable variation with a span between eight and 88 unique third party servers. The current study wants to find out 1) if these comparisons between various types of legacy broadcast media hold and 2) whether GDPR has affected the presence of third-party servers.

Other studies have focused on understanding the ownership and type of these third-party servers. Lindskow (2016) mappeds the business network of 41 US media publishers and finds 1356 business partners involved in building web pages for users, concluding that traditional news media webpage production involves a huge network of interacting companies. Kammer's (forthcoming) analysis of the use of third-party trackers by a sample of 25 news apps of legacy media suggests a divergent yet wide range of trackers as well as an increase in extent and complexity of the network over time but also shows that behind this wide variety of tracking services are a smaller group of dominant players, such as Alphabet. Our study identifies, maps and analyses the third-party servers on commercial broadcast media and PSM websites , with a focus on advertising, before and after GDPR. To understand the relevance of our research questions, we build a theoretical framework around PSM, third-party servers, privacy and trust.

Protecting privacy: from the legal to the ethical

The privacy concerns involved in the use of third-party trackers has received considerable attention from law and policy makers, looking to protect personal identifiable information (pii) and privacy in general. National law and policy makers, however, are hampered in their response options on account of the dominance of neo-liberal and free speech paradigms (Van Dijck, 2014), a lack of technological understanding of the issues at hand (Van den Bulck et al., 2018a), and, most of all, limited options to tackle global companies that escape the grasp of any national control (Vaidhyanathan, 2018). Instead, the EU has taken the lead, defining requirements for providers of interactive services regarding personal data protection and obtaining users' consent about data collection (EU Parliament, 2002; The EU Internet Handbook, 2016). However, abovementioned cookie-less 'fingerprinting' technologies were not mentioned in the 2002 EU regulation (Directive 2002/58/EC) resulting in the EU developing its General Data Protection Regulation (GDPR), in force since 25 May 2018.

GDPR provides extended rights to users to protect personal information, including the right to be informed about the processing of pii, and the right-to-be-forgotten, i.e. have all pii removed from the provider's records.

However, collecting user data involves more fundamental ethical questions. Indeed, data collection can be lawful but can still be at odds with ethic principles of good behaviour towards end users. A key issue in this revolves around who exactly is expected to be ethical? Is it restricted to third party tracker services themselves or also the media that use them? Much discussion recently revolves around the ethical role of tech and social media giants such as Facebook, as the size and power of these companies escapes control. A good illustration hereof is the 2019 FTC fine of $5 billion for Facebook's violation of various privacy rules. While this amounts to the highest fine in FTC's history, the number is dwarfed by Facebook's $15 billion revenue in the Spring quarter of 2019 and its $22 billion profit in 2018. In a cynical turn of events, the fine resulted in Facebook's stock prices going up (Patel, 2019). This relative inability, in most legal contexts combined with a strong reluctance, to regulate and curb these Tech Giants, has resulted in law and policymakers appealing to these companies' ethical awareness. As Wagner (2018:1) suggests, though, in this context, '"ethics" is the new "industry self-regulation"' with government regulation considered as part of the problem rather than the solution and Tech Giants turning to ethics as a catch-all phrase and a means 'to be seen to be doing it'.

Far less discussed, though, is the ethical behaviour of legacy media, both private and PSM. Indeed, data collection may be lawful but can still be at odds with the responsibility of legacy media towards its users, especially, PSM institutions that are considered and have an obligation as so-called 'islands of trust'. This has two complementary theoretical perspectives: media values, especially PSM values, and computer ethics. The latter offers analytical tools to expand analysis of data collection beyond privacy to include trust and security. Moor (1997) considers digital data as 'greased information': once information becomes a digital signal, it becomes unstoppable, beyond the grasp of an individual, organization or business. This has implications for privacy and security. For authors like

Moor (1997), privacy is not a legal or philosophical issue in its own right but is instrumental to the core human value of security. For individuals and societies to function, not everybody should know everything. As such, Moore worries more about the potential than the actual harm that greased information poses to personal security. In contrast, authors like Thompson (2001) see privacy as a risk function of exposing information: for each item of information disclosed, potential privacy consequences can be estimated.

Data collection and analysis also have ethical implications for the individual in relationship to society, i.e. to individual agency and freedom. Brey (2005) discusses how decisions taken by algorithms based on user data may not empower the user, but rather, remove agency by making assumptions, by not providing opportunities to correct wrong assumptions and by serving companies instead of user interests. Similarly, Vedder (1999) argues that algorithmic systems that use aggregated data covering many users, like a credit-scoring system for bank loans, de-individualizes citizens. Without individual assessment, a decision cannot be fair. Bozdag (2013) further shows that algorithmic recommendation systems, e.g. for media content, are biased, i.e. reflect specific ethical values. These issues raised by computer ethics are important to our research question, since user data collection and analysis form the basis for 'calculated public spheres' (Birkbak and Carlsen, 2016; Harper, 2016), extending the issue to the relationship between legacy broadcast media and algorithms (Van den Bulck & Moe, 2018).

(Legacy) Media, Trust and the PSM Conundrum

The issue is also relevant from the point of view of key values of (legacy) media and of PSM in particular. Traditional, media have been considered as cornerstones of democracy through the creation of a public sphere, the fourth estate watchdog of government and corporate power. As a result, the media were a place of trust, where people could turn to in the conviction that these media had their best interests as citizens at heart, even in commercial media systems and with varying journalistic and entertainment values. In return, and for a very long time, at least in the US and Europe, people showed great trust in 'the media'. While a slow decline has been going on for over a decade, more recently this rust can no longer be taken for granted, following a range of trends including tabloidization, personalization and recent developments such as the upsearch in dis- and misinformation (Jones, 2018, Stoll, 2019). Issues relating to privacy could further undermine this trust, a key reason for survival in a media market in decline (van der Burg & Van den Bulck, 2017).

These issues are of particular relevance to PSM as they differ from other media in that governments and societies expect a commitment to ensure universality, contribute to identity and social cohesion, and provide a mix of information, inspiration and entertainment, while maintaining high levels of trustworthiness (Van den Bulck & Moe, 2018). Always intrinsic but implicit in PSM ideals, from the 1990s onwards, trust became an explicit topic in legitimating its relevance as PSM institutions' fought fierce commercial competition and 'hostile' governments. The European Broadcasting Union's (EBU) - PSM institutions representative and lobby organization - Digital Strategy Group (2002) identified the PSM institution as an 'island of trust' amidst an increasingly commercial and self-serving ecosystem (Bardoel and d'Haenens, 2008). Biltereyst (2004: 341) calls it the 'aura of trust' that 'includes a feeling of quality, reliability, honesty, competence and good intentions'. Research confirms that in countries with strong PSM, trust in radio and television broadcasting is stronger (EBU, 2017). At the same time, to remain relevant, PSM institutions, too, need to engage with digital developments, including the use of third party trackers, to remain relevant in a media-ecosystem increasingly dominated by digital platforms and personalized services and, thus, influenced by algorithms and big data. While this helps to better serve the citizen/user, it can contribute to an economy of user data that may be of little benefit to the institution or the citizens, and that may challenge PSM's core values. This tension reverberates the classic challenge for PSM, suspended between reaching audiences with interesting content in a crowded environment and providing 'added public value' to users-as-citizens, away from commercial pressures.

METHODOLOGICAL SET-UP

To understand the use of third party trackers by private and public service media before and after GDPR, to analyse relevant differences between media types and to get a better understanding of the types and ownership of these services, we followed a procedure tested in previous research (Sørensen & Van den Bulck, 2018). As such, our data result from an extensive and repeated collecting of third party traffic on media-related websites. From a dataset of +32 million recordings of HTTP responses from servers for files like pictures, code or text to +12700 web pages from 1250 websites visited 9 times before (from 2018-02-05 to 2018-05-05) and 24 times after GDPR (from 2018-05-25 to 2019-04-12), we selected 348 media websites from 39 European countries (#113 from EBU members, #235 from private media), see (Sørensen and Kosta, 2019). We analysed the presence of third party URLs (in the following 'TP URLs') in our browser at the level of HTTP responses, including the many different page elements, e.g. scripts, pictures, fonts, videos that are used to build a webpage in the user's browser. We can thus determine the percentage of page elements that are delivered from other sites ('third-party servers') than the media website that we visited. Furthermore, for each site we visited, we can count how many different TP URLs we meet while browsing the site. In order to imitate normal user's browsing behaviour on a media site, we have visited 10 randomly chosen pages from each media site. The visited pages remained the same throughout all our 33 visits. We found 3256 unique TP URLs, for 1517 of which we could identify the company.

As TP URLs serve many different purposes, such as content delivery, analytics, advertising, amongst others (see above), we manually visiting each TP URL that appears at more than one of the visited sites and evaluated the purpose of the third party interaction by reading third party website descriptions of the services offered by the third party company. This was supplemented by looking up identified TP URLs in databases, such as WhoIs (https://www.whois.com, accessed 2019-06-01) and Threatcrowd.org, accessed 2019-06-15). Coding the purpose of the TP URLs in a multi-stage coding process, we defined 16 categories: *Advertising, Analytics, Content, Cybersecurity, Distribution technology, Editorial, Malicious, Plug-in, Privacy, Programming, Publisher, Retail, Search engine, Social media,* *Unidentifiable and Specific to a certain media site - not categorized (see appendix 1).* URLs that only appear at one site has not been categorized.

Analysing the media sites, one main distinction we used is between privately owned media websites, and media websites offered by PSM organisations (in this case, all members of the European Broadcasting Union (EBU). Although we have a few media websites from countries outside Europe, in this analysis we focus on media websites from European countries (either in- or outside EU/EEA). Media websites were categorized in two further ways to find patterns that cut across the private/public distinction. One, is to look at the number of unique TP URLs found at each site in relation to the ratio of page elements from TP servers. In this way we get a plot with x- and y- axis. We have indexed the number of unique TP URLs so the site with the highest number (bfmtv.com from France) is represented with 1, and all other sites a fraction of 1. This plot can be seen in Figure 01:

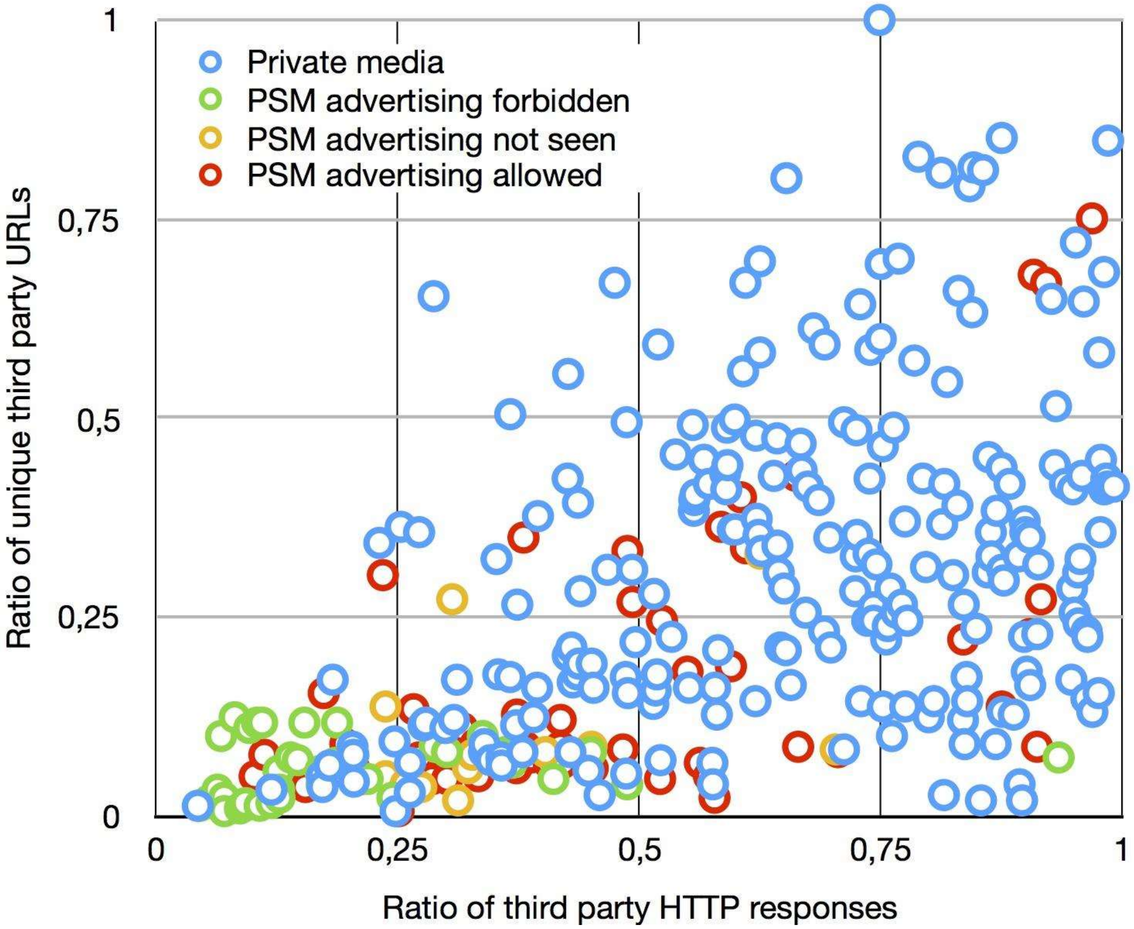

Figure 1: # unique TP URLs found at each site in relation to the ratio of page elements from TP servers for four types of media,

Beside that, we look at the distribution of TP URLs across the above-mentioned categories of TP services. To identify media sites that have similar patterns in the distribution across the different types of TP services, we used a clustering software ('Rapidminer') to identify clusters of media websites. To prepare values between 0 and 1, needed for the clustering, the number of TP URLs for each site and TP-category were divided by the total number of found TP URLs. Subsequently, using X-means and Euclidean distance for the calculation, four clusters were found. In the analysis we will discuss the characteristics of these clusters.

As our measurements span over a long period of time that includes an event (introduction GDPR) assumed to impact on third party activity on webpages, we present diagrams that depict developments over time. The many dimensions of the data and calculations allow for a wide range of possible analysis and visualisations, only a few of which we can discuss in the scope of this contribution.

RESULTS / ANALYSIS

Occurrences and distribution of TP URLs across various categorisations

Across the 33 visits, we found 3256 unique TP URLs, spread into 2944 unique TP URLs for the 235 private media sites and 1013 for the 113 public media sites. Analysing over time, Figure 2 shows a dramatic drop in number of unique TP URLs between May 5, 2018 and May 29, 2018 for private media, while PSM sites show a slower decline.

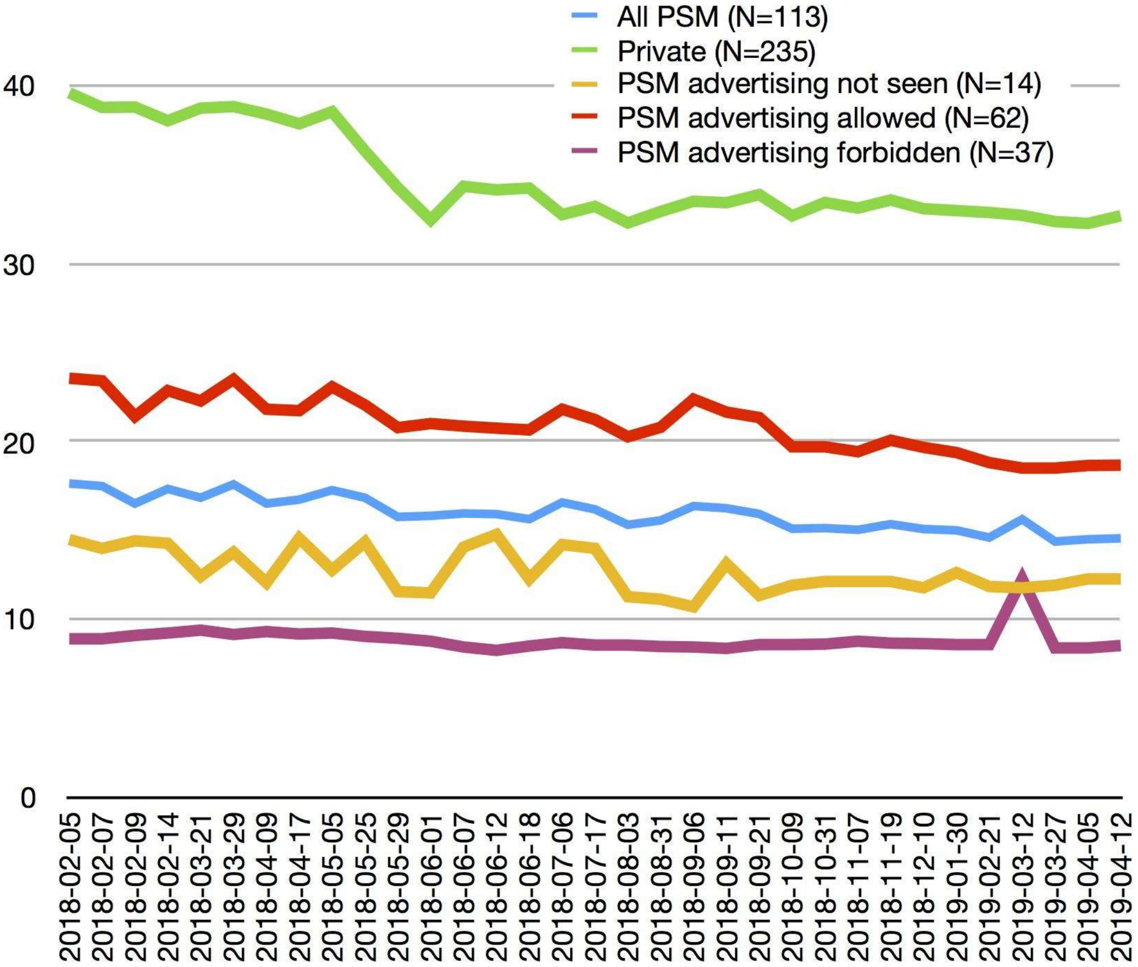

Figure 2: evolution of TP URLs over time for four types of media

PSM where advertising is forbidden have the lowest level of unique TP URLs per site and visit, confirming earlier research results (Sorensen & Van den Bulck, 2018). Furthermore, it is clear that fluctuations occur for all types of websites, reflecting the dynamic nature of webpage production characterised by rapid introduction of new technologies and change of suppliers of web services, amongst others.

Beyond that, results presented in Figure 3 show that the number of different TP URLs for each of our visits declined over time.

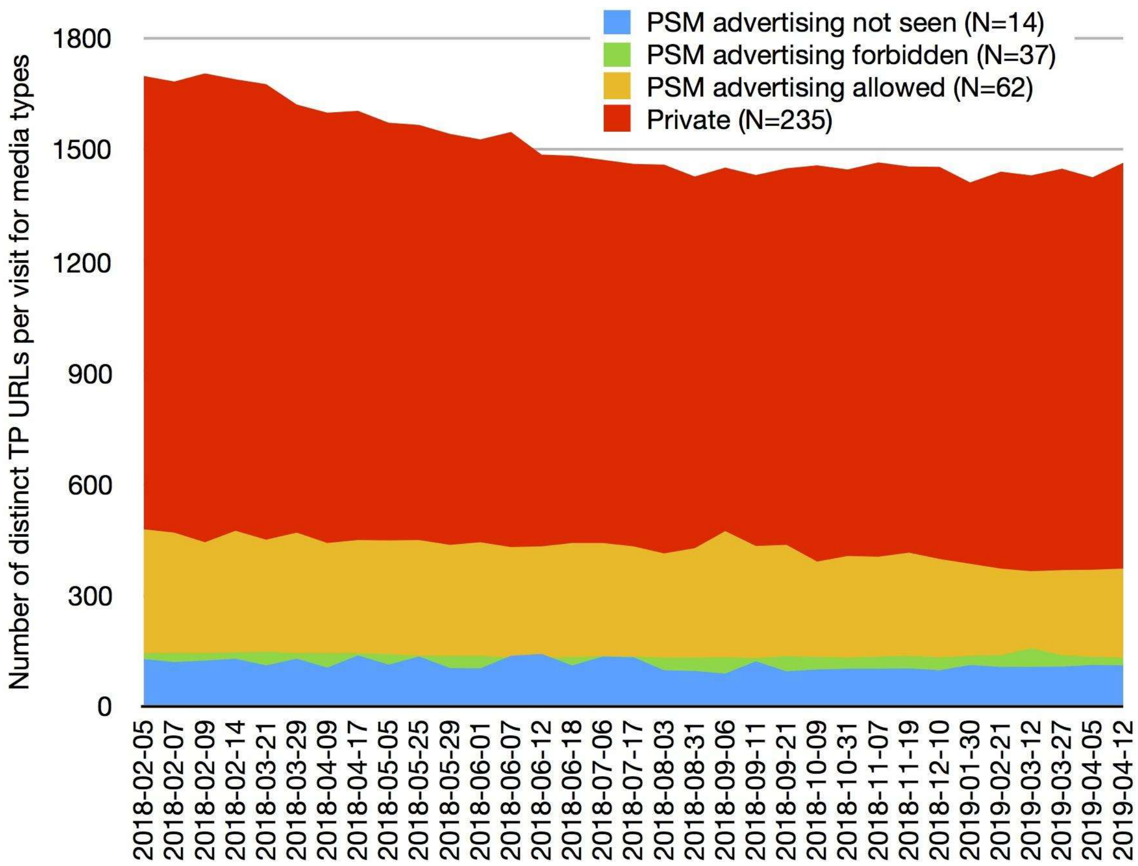

Figure 3: number of distinct TP URLs over time for four types of media

For PSM sites where advertising is either forbidden or not seen by us, the total number of different TP URLs fluctuate between 88 to 141 (PSM advertising not seen) and 130 to 156 (PSM advertising forbidden), although the overall decline is negligible.

In Figure 4, we compare our first nine measurements / visits to media sites, which took place before May 25, 2018, with the last nine measurements / visits (November 7, 2018 to April 12, 2019).

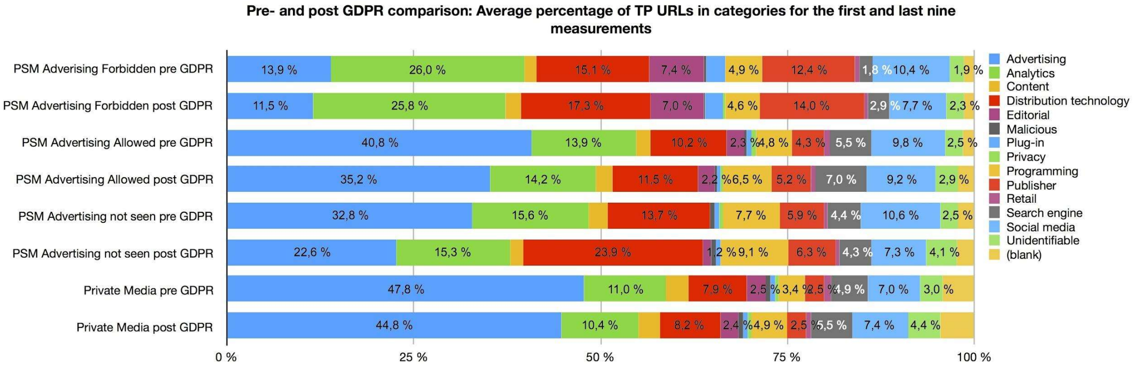

Figure 4: average % of TP URLs by type of servers pre and post GDPR

Figure 4 shows the share of the different TP URLs categories in two time-intervals respectively before and after GDPR. Marked decrease in occurrences can be observed for Advertising TP URLs and, to a lesser extent, for Analytics and Social Media (except on Private Media sites). Conversely, a marked increase can be observed for the category of Distribution technology for PSM sites where no advertising was observed. In volume, the average number of TP URLs declines for most categories of TP URLs and on most types of media sites in the nine post GDPR measurements. Growth appears however for some TP URLs categories on some types of sites: For PSM sites not allowed to carry advertising: Content (32%), Distribution technology (13%), Publisher (11%), Search engine (56%) and Unidentifiable (+22%); for PSM sites allowed to carry advertising: Programming (14%), Publisher (2%), Search engine (6%) and TP URLs that only occur on one site (31%); for PSM site where no advertising was observed: Distribution technology (+54%), Programming (+4%) and Unidentifiable (+48%); and, finally, for private media: Programming (+21%) and Unidentifiable (+24%). Figure 5 provides further details.

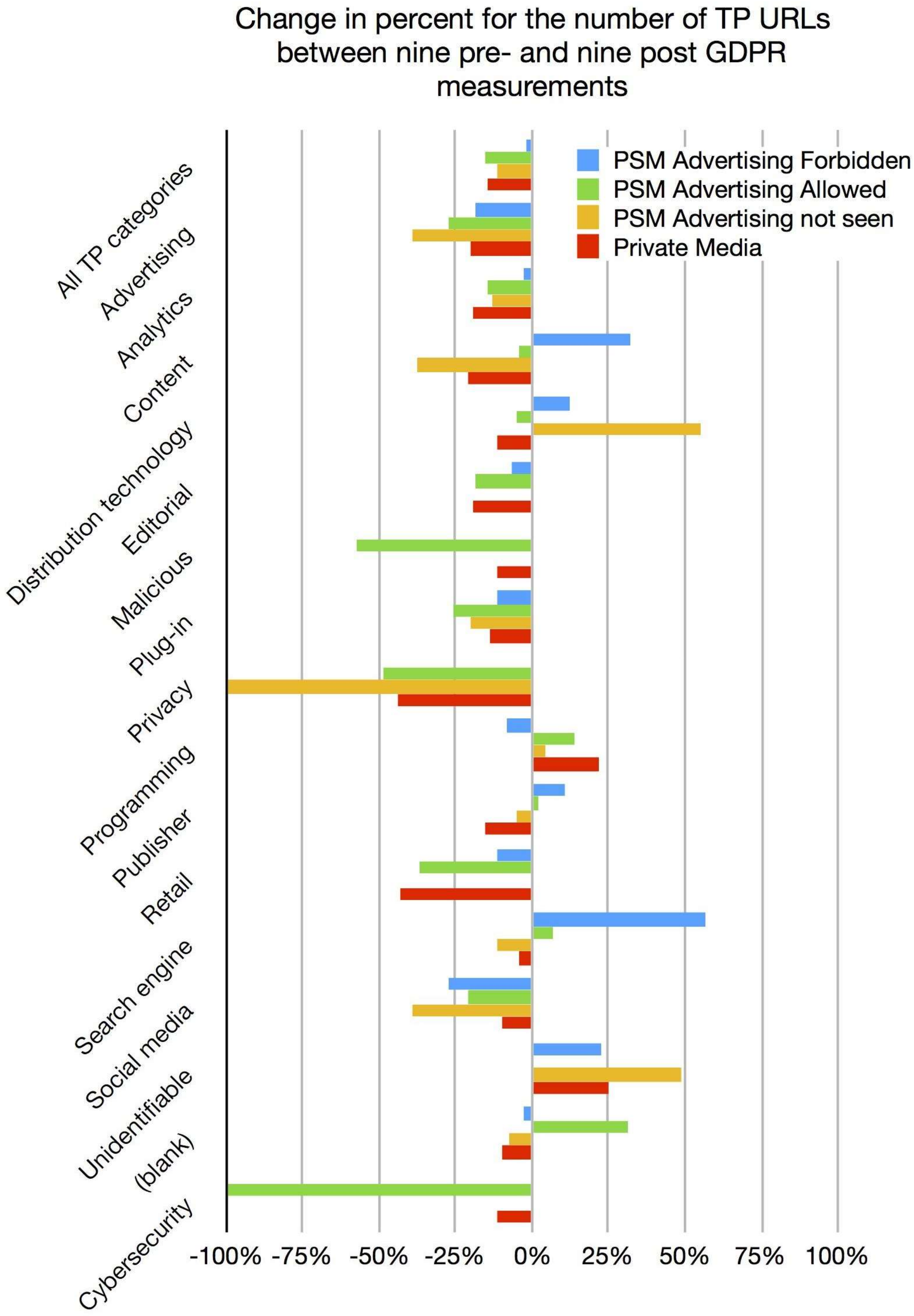

Figure 5: Evolution in types of TP URLs before and after GDPR by four types of media

If we look at the plot of the sites in relation to their use of external servers to produce the webpage (the HTTP-ratio) and their indexed number of TP URLs, depicted in figure 1 (see earlier), we see that most PSM sites that are not allowed to carry advertising both have a low number of unique TP URLs and produce the web pages with own servers. these PSM sites have a lot in common.

Conversely, many private media have a high ratio of unique TP URLs and produce their web pages using many different third party servers. Interesting, sites from PSM that are allowed to carry advertising (circled in red), to a certain degree, follow the patterns of private media rather than that of the other PSM sites. Furthermore, the plot shows a diversity within private media sites. Upper right corner contains sites with a relatively high number of different third party URLs and a high ratio of page elements (represented by HTTP responses) from third party servers.

Media Clustering beyond the Public-Private division

Using data mining software we aimed at identifying groups of media sites not defined by being private or public, but by the characteristics of the type of TP URLs that users meet when visiting the media pages. Running an X-means clustering algorithm on data of the number of TP URLs for each site distributed over the 16 categories of TP URLs purposes (see above) the software revealed four clusters, as presented in Table 1.

	No of sites	Private media	All PSM sites	PSM Advertising forbidden	PSM Advertising not seen	PSM Advertising allowed
Cluster 0	115	79	36	5	3	28
Cluster 1	86	26	60	31	9	20

Cluster 2	41	38	3	0	0	3
Cluster 3	101	88	13	1	1	11

Table 1: X-means clusters based on # TP URLs for each site distributed over the 16 categories of TP URLs purposes

Cluster 0 is mainly (68%) composed of Private sites, mostly (72%) from the EU, while it also has the second highest share of non-EU sites. Advertising TP URLs comprise 36% of all TP URLs, followed by 14% Analytics, Social Media 11%. This cluster has 1247 unique TP URLs of which 458 Advertising-related, 76 Analytics-related, 20 Malicious and 108 Unidentifiable.

Cluster 1 is composed mainly (70%) of PSM sites, with slightly more sites where advertising is forbidden (31) than possible (20). In this cluster, 86% of sites are from the EU. Advertising TP URLs comprise 28% of all TP URLs, Analytics 20%, Distribution technology and Publisher each 11%. The Cluster presented 611 unique TP URLs, of which 201 Advertising-related, 8 Malicious and 35 Unidentifiable.

Cluster 2 is clearly dominated (92%) by private media with only 3 PSM sites all on which advertising is allowed. Cluster 2 has, with 24%, the highest share of non-EU sites. More than 50% of the third party URLs in Cluster 2 are Advertising-related, while Analytics constitute 8% and Distribution technology 9%. Cluster 2 has 1647 unique TP URLs, of which 595 are Advertising-related, 37 Malicious and 117 Unidentifiable.

Cluster 3 is dominated by private media (87%), accompanied by 12% PSM sites, of which 11 are PSM where advertising is allowed, and one each of the other types of PSM sites. Cluster 3 has the lowest share of non-EU sites. Advertising TP URLs comprise 47% of all TP URLs, Analytics 11%, Distribution technology 9%. The cluster has 1727 unique TP URLs, of which 665 are Advertising-related, 22 Malicious and 142 Unidentifiable.

Analysing the presence of TP categories in the different clusters, Cluster 1 stands out as the one with the lowest number of unique advertising-related TPs and very few malicious and unidentifiable TPs. It also has in general the lowest number of different TP URLs, which can produce the preliminary conclusion that users that visit sites belonging to Cluster 1 expose their data privacy to a lesser degree than in the case of the other clusters. The distribution of third parties in categories is provided in Figure 6

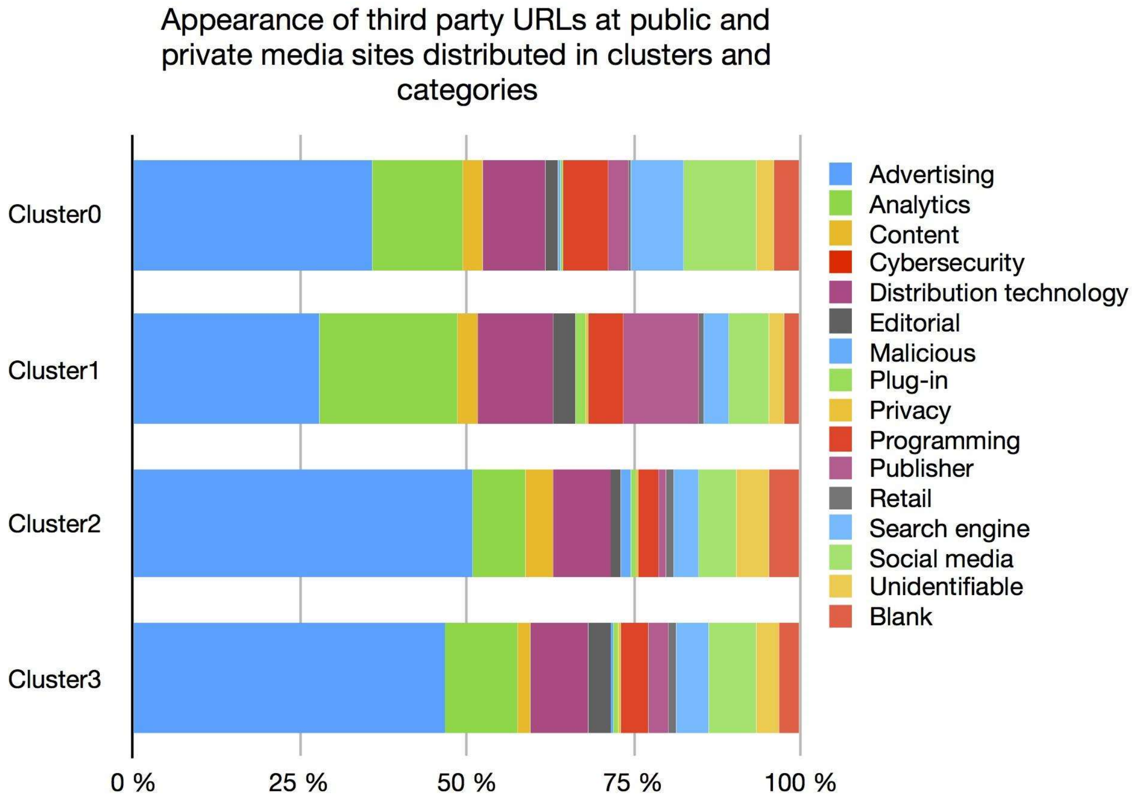

Figure 6: Occurence (in %) of types of third party servers by cluster

Companies in the clusters

FInally, we want to shed some light on the main players in this field of third party services. In light of wider developments in digital media and technology, it is not very surprising that these services are dominated by the big tech giants. TP URLs from Google are present at 31% of the pages, from Facebook at 7% of pages. The Top-20 that excludes Google and Facebook shows the major third party companies are not equally well represented in all four clusters. Cluster1, characterized by a high number of PSM sites, is interesting as Comscore (analytics), Chartbeat (editorial), New Relic (analytics) and - to a smaller extent - Twitter and hotjar (analytics) are more present in this than in the other clusters. Figure 7 presents the top 20 third party companies.

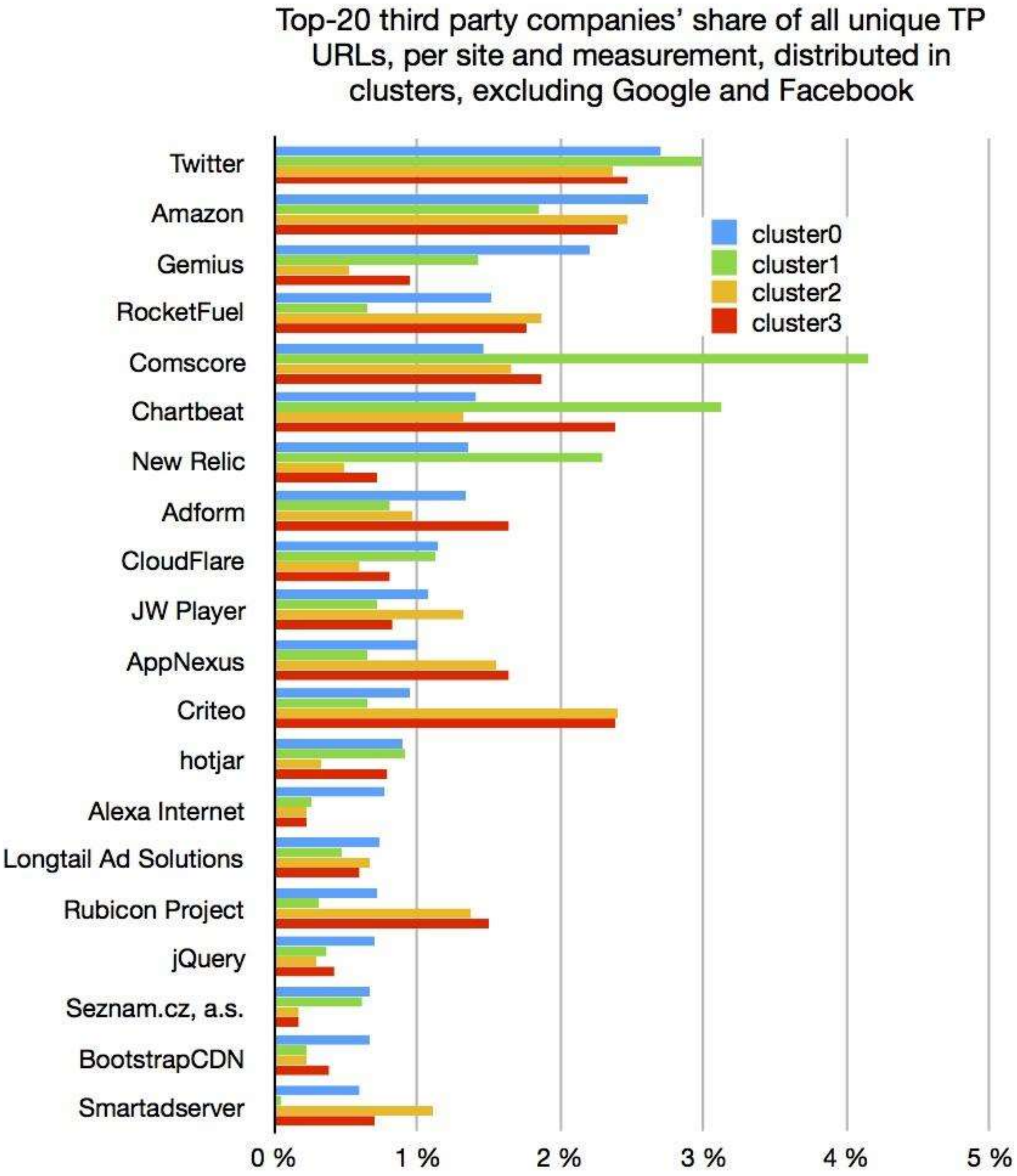

figure 7: Top 20 Companies of third party services by cluster

DISCUSSION AND CONCLUSION

Our data show that media websites include scripts that contact third-party servers, albeit to different degrees. Before drawing conclusions, we note that the number of different third-party URLs must be interpreted with care. For one, the same third-party company can be present through several server names, which can account for unidentifiable and technical third-party servers. Conversely, Englehardt and Narayanan (2016) and Lindskow (2016) show that the number of new third-party servers increase with additional visits so while we visited the sites repeatedly before and after GDPR, results are still influenced by the number of iterations. Furthermore, not all third parties are equally influential or important in the question of pii user data, so a particular website may have a low number of third party URLs but these can be very influential third-party services like Google or Facebook. Moreover, if a third party is present on several web pages, the description of the user is more precise and, thus, more valuable to advertisers. Overall, the size of our data may be affected by the number of visits, but our data represent all possible third parties a user can encounter. Indeed, our analysis shows that many third-party servers are programmatic, i.e. triggered by embedded scripts that depend on a user's browser history (cookies), device history (fingerprinting), location (geo-location), match with existing user-profile data e.g. from social networks, or wrapped scripts in scripts.

When analysing the PSM websites, we see certain public service media organisations keep the involvement of third-party servers at a very low level. However, other PSM media - particularly those that show advertising on their websites or are allowed to do so - have a much higher level of different third party servers, and these also in many cases play a bigger role for composing the webpage. There is however a good explanation to the larger numbers of third party servers at pages with advertising - the process of selling advertising to advertisers involves a number of servers, just to find the right bid and buyer for the advertisement. However, as the bidding technology used for the sale of online advertising is currently at many websites being replaced with a system where bidden takes place not in the user’s computer, but between the media server and the advertising servers (so-called ‘server-side header bidding’ - see IAB, 2018a), the general decline we see in the number of third party servers may not necessarily reflect a lower exposure of user data, just that we cannot measure it any longer. That said, our impression is that GDPR led media publishers and advertising technology companies to clean up some unused servers and scripts and thereby also reduce the exposure of user data. But as we see when we look at development in other categories of third party servers, the general tendency goes in the direction of media websites to a greater extent using external web services for delivering the content and analysing the user behavior.

GDPR has initiated a process of regulation that has resulted in a formalisation between the web partners involved in the production of the media web pages. However, it is too early to conclude that GDPR has led to less exposure of user data. Rather GDPR may have resulted in a concentration of fewer third party service providers, with a few very strong ones among those that have gained from GDPR. Our clustering analysis shows that the media websites are heterogeneous in their use of third party services.

Our results illustrate what we consider to be a PSM dilemma. Our data confirm Lindskow’s observation (2016; 2017) regarding news webpages in the US and Denmark for a wide sample of PSM. Many PSM organizations are clearly deeply integrated in international networks when delivering their webpages, interacting with an extensive network of digital business partners that aggregate content, analyse user behaviour, sell or buy advertisements, integrate social media or simply deliver files and scripts. This helps PSM organizations to optimize editorial work and (where allowed) advertising revenue, and to develop personalized recommendations for its users. It further allows PSM to keep up-to-date with the newest technologies, platforms and user interfaces, and to reduce the need for PSM to invest in technology. The introduction of GDPR affected the number of these third party trackers to a relatively limited extend, suggesting that the introduction of such a legal framework does not change the core issue of media users being tracked. This comes at the price of dependency on (commercial) outsiders in content production and dissemination. PSM thus find themselves caught up in a dilemma between maintaining their integrity or participating in the exposure economy increasingly managed by international companies. Google and Facebook may be the best-known examples of the latter, but our research shows that they are just the most visible among hundreds of companies in the business of user data.

This dilemma can be considered as an ethical issue for PSM and policy makers: Can PSM organizations use the same tools as commercial media as freely as commercial media to monitor and optimize attention - tools that operate in the background without the knowledge of the user? Some arguments in favour include the need for PSM to be competitive with commercial media, to maintain relevance for users and to produce value for licence-free/public funding. However, as trusted institutions, PSM organizations have an ethical obligation to be honest and transparent in their mode of operation. If nothing else, opaque use of third-party servers undermines their very role as 'islands of trust', an important legitimation of their funding and existence.

REFERENCES

Acar G, Eubank C, Englehardt S, et al. (2014). The Web Never Forgets. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, New York, New York, USA, 2014, pp. 674–689. ACM Press. DOI: 10.1145/2660267.2660347.

Acquisti, A.; Taylor, C. & Wagma, L. (2016). The Economics of Privacy, *Journal of Economic Literature*, 54 (2): 442-492.

Bardoel J and d'Haenens L (2008). Reinventing public service broadcasting in Europe: Prospects, promises and problems. *Media, Culture & Society*. 30(3): 337–355. DOI: 10.1177/0163443708088791

Biltereyst D (2004). Public service broadcasting, popular entertainment and the construction of trust. *European Journal of Cultural Studies*. 7(3): 341–362. DOI: 10.1177/1367549404044787

Birkbak A and Carlsen HB (2016). The public and its algorithms: comparing and experimenting with calculated publics. In: Amoore L and Piotukh V (eds) *Algorithmic life : calculative devices in the age of big data*. Routledge, pp. 21–34.

Bozdag E (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3): 209–227. DOI: 10.1007/s10676-013-9321-6.

Brey, P. (2005). Freedom and Privacy in Ambient Intelligence. *Ethics and Information Technology*, *7*(3), 157–166. Doi: 10.1007/s10676-006-0005-3

Busch, O. (Ed.). (2018). *Programmatic Advertising*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-25023-6

Dolata, U. (2017). Apple, Amazon, Google, Facebook, Microsoft: Market concentration - competition – innovation strategies. *Stuttgarter Beiträge zur Organisations- und Innovationsforschung, SOI Discussion Paper, No. 2017-01*, Institut für Sozialwissenschaften, Universität Stuttgart, Stuttgart

Donders, K. Enli, G.; Raats, T. & Syertsen, T. (2018). Digitisation, internationalisation, and changing business models in local media markets: an analysis of commercial media's perceptions on challenges ahead, *Journal of Media Business Studies*, 15 (2). Online first doi.org/10.1080/16522354.2018.1470960

Donders, K. & Van den Bulck, H. (2016). 'Decline and fall of public service media values in the international content acquisition market: An analysis of small public broadcasters acquiring BBC Worldwide content, *European Journal of Communication*, 31(3): 299-316. DOI: 10.1177/0267323116635833

EBU Media Intelligence Service (2017). *Market Insights: Trust in Media*.

EU Parliament (2002). *Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)*. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:HTML

IAB Europe. (2016). *ADEX BENCHMARK 2015 EUROPEAN ONLINE ADVERTISING EXPENDITURE*. Retrieved from https://www.iabeurope.eu/wp-content/uploads/2016/07/IAB-Europe_AdEx-Benchmark-2015-report_July-2016-V3.pdf 2019-07-22

IAB - Interactive advertising bureau Europe. (2018a). *Header Bidding and Auction Dynamics*. Retrieved from https://www.iab.it/wp-content/uploads/2018/09/IAB-Europe_Header-Bidding-and-Auction-Dynamics-White-Paper_August-2018-1-compressed.pdf

IAB Europe. (2018b). *AdEx Benchmark Study 2018*. Retrieved from https://www.iabeurope.eu/wp-content/uploads/2019/06/IAB-Europe_AdEx-Benchmark-FY-2018-study_website_FINAL.pdf 2019-07-22

IAB (2019). European Digital Advertising Market Exceeds €55bn in 2018. Retrieved from https://www.iabeurope.eu/all-news/press-releases/european-digital-advertising-market-exceeds-e55bn-in-2018/

Falahrastegar M, Haddadi H, Uhlig S, et al. (2014). *Anatomy of the Third-Party Web Tracking Ecosystem*. Retrieved from: http://arxiv.org/abs/1409.1066.

Harper, T. (2016). The big data public and its problems: Big data and the structural transformation of the public sphere. *New Media & Society*, Doi: 1461444816642167.

Internet Engineering Task Force (2011) RFC 6265: HTTP State Management Mechanism (‘HTTP cookie’). Available at: https://tools.ietf.org/html/rfc6265 (accessed 2 November 2017).

Jones, J.M. (2018). U.S. media trust continues to recover from 2016 low, *Gallup*, October 12,

Retrieved from https://news.gallup.com/poll/243665/media-trust-continues-recover-2016-low.aspx

Kammer, A. (forthcoming). Resources exchange and data flows between news apps and third party actors: The digital transformation of the news industry, *New Media and Society*, Lindskow K (2016). *Exploring digital News Publishing Business Models - A Production Network Approach*. Copenhagen Business School. Available at: http://hdl.handle.net/10398/9284.

Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique, *Big Data and Society*, July-Sept, DOI: 10.1177/2053951714541861

Mayer-Schoenberger, V. and K. Cukier (2013). *Big Data. A revolution that will transform how we live, work, and think*. London: John Murray Publishers

Mayer-Schoenberger, V. & Padova, Y (2016). Regime change? Enabling Big Data through Europe's new data protection regulation, *The Colombia Science and Technology Review*, 17 (Spring).

Moor, J. H. (1997). Towards a theory of privacy in the information age. *ACM SIGCAS Computers and Society*, *27*(3), 27–32. Doi: 10.1145/270858.270866

Patel, N. (2019). Facebook's $5M FTC is an embarrassing joke: Facebook gets away with it again, *The Verge*, July 12, retrieved from https://www.theverge.com/2019/7/12/20692524/facebook-five-billion-ftc-fine-embarrassing-joke

Raats, T., Evens, T. & Pauwels, C. (2015). Towards sustainable financing models for television production? Challenges for audiovisual policy support in small media markets. Proceedings of the 30th European Communications Policy Research Conference (EuroCPR) 2015, March 23-24, 2015, Brussels, Belgium.

Reseke, L. (2018). JP/Politiken tager opgør med annonce-jungle og dropper 200 samarbejder. https://mediawatch.dk/secure/Medienyt/Web/article10631483.ece

Srnicek, N. (2017). *Platform Capitalism.* London: Wiley

Stoll, J. (2019). Trust in media in Europe: Statistics and facts, *Statista*, March 18th, retrieved from https://www.statista.com/topics/3303/trust-in-media-in-europe/

Sørensen, J. K., & Kosta, S. (2019). Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *WWW '19 Companion*

Proceedings of the The Web Conference 2019. ACM. https://doi.org/10.1145/3308558.3313524

Sørensen, J.K. & Van den Bulck, H. (2018). Public service media online, Advertising and the third-party user data business: A trade versus trust dilemma?', *Convergence, The International Journal of Research into New Media Technologies*, Online first: 1–25

The EU Internet Handbook (2016). Cookies - European Commission. Available at: http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm

Thompson PB (2001). Privacy, secrecy and security. *Ethics and Information Technology* 3(1): 13–19. DOI: 10.1023/A:1011423705643.

Turow, J. (2011). *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. Yale University Press.

Vaidhyanathan, S. (2018). *Anti-Social Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford: Oxford University Press.

Van den Bulck, H. Donders, K. & Lowe, G.F. (2018a). Public service media in the networked society: What society? What network? What role?, pp 11-28 in G.F. Lowe, H. Van den Bulck & K. Donders (Eds.) *Public Service Media in the Networked Society*. Gothenburg: NORDICOM.

Van den Bulck, H. & Moe, M. (2018). Universality and personalisation through algorithms: Mapping strategies and exploring dilemmas', *Media Culture and Society,* 40(6): 875-892.

van der Burg, M. & Van den Bulck, H. (2017). Why are traditional newspaper publishers still surviving in the digital era? The impact of long-term trends on the Flemish newspaper industry's financing, 1990–2014, *Journal of Media Business Studies*, 14 (2): 82-115. DOI: 10.1080/16522354.2017.1290024

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology, *Surveillance and Society*, 12(2) https://doi.org/10.24908/ss.v12i2.4776

Vedder A (1999). KDD: The Challenge to Individualism. *Ethics and Information Technology* 1(4): 275–281. DOI: 10.1023/A:1010016102284.

Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), *Being Profiling. Cogitas ergo sum*. Amsterdam University Press.

Wambach T and Bräunlich K (2017). The Evolution of Third-Party Web Tracking. In: *Information Systems Security and Privacy. ICISSP 2016. Communications in Computer and Information Science*, 2017, pp. 130–147. Springer. DOI: 10.1007/978-3-319-54433-5_8

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York, NY: Public Affairs

Appendix 1 TP URLs typology

The Advertising category encompases services related to the sale of ads and analysis of user profiles. From descriptions on websites presenting third-party services we have included Programmatic advertising services, Personalization services, Recommender systems, Real-time bidding platforms, Demand-side and Sell- side platforms, Data management platforms and data inte- gration, 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 advertising services, Cross-device user identification, and Video-based advertising in the category.

Analytics contains 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), Customerflow, 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 contains services that perform internet infrastructure surveillance.

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

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

Malicious are servers / URLs that could not be identified, but when examined in the cybersecurity community Threatcrowd - https://www.threatcrowd.org were voted as ‘Malicious’ by the users.

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 scripts, fonts and other tools rendering the webpage.

Publisher are servers that are owned by visited 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 are URLs that do not return a readable HTML page, but a 404 message, a blank page, a time-out or an access forbidden message.