The Impact of the GDPR on Content Providers

V. Lefrere,[*]L. Warberg,[†]C. Cheyre,[‡]V. Marotta,[§]and A. Acquisti[¶]

June 2020

PRELIMINARY WORKSHOP DRAFT - WORK IN PROGRESS
PLEASE DO NOT CITE
CONTACT AUTHORS FOR LATEST VERSION

Abstract

We study the impact of the European General Data Protection Regulation (GDPR) on the advertising-supported online ecosystem. We focus on online content providers (such as news websites) and their users. We investigate whether restrictions on online tracking enforced by the regulation ultimately affect downstream variables such as the quantity of content that websites offer to their visitors and users' engagement with such content. The results suggest that the GDPR reduced the number of third-party cookies and tracking responses in both US and EU websites. Furthermore, the enactment of the GDPR may have to some extent negatively affected traffic to EU websites, relative to US websites. However, the enactment does not seem to have negatively affected the amount of content that EU websites were able to publish (relative to US websites), or the degree of average social media engagement and interaction with such content. Our analysis is ongoing, as data collection is continuing.

[*]Institut Mines Telecom, Business School. Email: vincent.lefrere@imt-bs.eu
[†]Engineering and Public Policy, Carnegie Mellon University. Email: warberg@cmu.edu
[‡]Department of Information Science, Cornell University. Email: ccheyre@infosci.cornell.edu
[§]University of Minnesota Twin Cities Carlson School of Management. Email:vmarotta@umn.edu
[¶]Heinz College, Carnegie Mellon University. Email:acquisti@cmu.edu.

1 Introduction

In May 2018, the European Union (EU) implemented the General Data Protection Regulation (GDPR) to increase the protection of users' privacy and individuals' control over their data. The enactment sparked interest among academics, policy-makers, and industry actors worldwide. Much focus has been devoted to measuring sites' compliance with the GDPR, documenting changes in online consent mechanisms (and whether such changes ultimately are effective in protecting users' privacy), and estimating compliance costs (See section 2). Less attention, so far, has been devoted to understanding downstream consequences that the regulation might have on economically important metrics, such as the ability of websites to produce content, and the ability of Internet users to access and benefit from it. Our ongoing study aims at contributing to the debate on the impact of the GDPR by investigating whether potential changes in the dynamics of online tracking brought about by the GDPR led to changes in the quantity of content that websites offer to their visitors and changes in users' engagement with such content.

A defining characteristic of the GDPR is the restrictions it places on the collection and processing of European (EU) residents' data by organizations. An organization may only process users' data if it satisfies one of several legal bases for data processing. One such basis is explicit opt-in consent from users. For example, when a user browses a website owned by a publisher, the publisher must request the user's explicit permission to allow cookies to be set on the user's machine and, if so, whether she would also allow tracking cookies by third parties (see Figure 8). The requirements for lawful processing should subsist even if the company operates outside the EU, as long as the data being collected or used belong to EU residents. This differs from the pre-GDPR *de facto* standards for most websites across the world. In absence of regulatory obligations, websites can track users' behaviors by default, often merely informing users that they implicitly consent to tracking by virtue of accessing the website. In addition to re-

quiring that organizations obtain consent for data collection and processing, the GDPR establishes steep financial penalties for organizations that do not comply. For example, the GDPR enabled the French privacy regulatory authority (CNIL) to fine Google €50 million for "lack of transparency, unsatisfactory information and lack of valid consent for the personalization of advertising."[1] This action, along with advice released by the UK Information Commissioner's Office, suggests that consent may be emerging as the primary basis for enabling data processing for behavioral advertising under the GDPR.[2] Explicit consent mechanisms and associated penalties for non-compliance can affect industries and markets which rely heavily on personal data and profiling. The online advertising industry is expected to be particularly affected by the GDPR since its growth is driven by the ability to track users' online behavior to deliver personalized advertising. In a recent study (which did not focus on GDPR, but on the impact of reductions in consumer tracking), Johnson *et al.* (2020) investigate an industry self-regulation initiative which allows American consumers to opt-out from tracking. They find that losing the ability to target advertising resulted in a decrease of around $8.58 of ad spending for each consumer who chose to opt out, beared by publishers and the exchange.

Given how widespread the collection and usage of data is across different sectors, the overall economic impact of the GDPR may be substantial. An early 2013 Deloitte impact assessment report suggested that the potential economic impact of the GDPR (Deloitte, 2013) could amount to a loss of around 2.8 million jobs and a reduction of European GDP by around 1.34% (corresponding to around €173 billion). Since its growth is driven by the ability to track users' online behavior to deliver personalized advertising (and explicit opt-in requests, in turn, are likely to interfere with such ability), the online advertising industry is expected to be particularly affected by the

[1] https://www.cnil.fr/fr/la-formation-restreinte-de-la-cnil-prononce-une-sanction-de-50-millions-deuros-lencontre-de-la

[2] https://ico.org.uk/media/about-the-ico/documents/2615156/adtech-real-time-bidding-report-201906.pdf

GDPR. First, the imposition of limitations on the ability to collect and use data could have a negative impact on the effectiveness of online advertising campaigns (Goldfarb and Tucker, 2011). Second, the GDPR may also impact the composition of the online advertising industry: regulation could impose comparably greater constraints and risks on small and medium-sized online advertising firms, leading to a further concentration of an industry where dominant players already have substantive power to define how the market operates and how benefits are allocated (Johnson and Shriver, 2019). This could in turn expose users, as well as firms upstream and downstream from advertising platforms, to other types of harms such as monopolistic behavior.

Reductions in advertising effectiveness, spending, and competition may ultimately negatively affect publishers. Since advertising is a major revenue component for digital goods producers (Lambrecht *et al.*, 2014), constraints on online tracking and consumer data gathering may ultimately threat the subsistence of free online content and services. Claims by both industry groups and think tanks support this hypothesis. In a late 2015 report by IHS Technology, the CEO of the Interactive Advertising Bureau of Europe, Townsend Feehan, suggested that overly burdensome privacy regulation may "limit digital advertising's ability to continue to deliver a wide range of online content to users at little or no cost at the point of consumption" (IHS Technology, 2015). An earlier report by the Information Technology and Innovation Foundation made stronger claims, stating, "The evidence clearly suggests that the tradeoffs of stronger privacy laws result in less free and low-cost content and more spam (i.e. unwanted ads) which is not in the interests of consumers" (Castro, 2010). Both sources capture the sentiment that overbearing privacy regulation could negatively impact publishers, resulting in a reduction in the availability of content or a degradation of its quality.

Despite the numerous claims and predictions about the potential effects of the GDPR on the profitability of ad-supported content providers and services, empirical evidence is limited and contradictory. For instance, anecdotal evidence suggests that at

least some online publishers that reduced their use of behaviorally targeted advertising in the EU, post-GDPR, have continued to enjoy stable advertising revenue (Davies, 2019). Additionally, it is unclear how consistently different companies interpreted and applied the regulation. Small firms could find implementing the GDPR difficult, and their size may not justify the investment necessary to use personal data in a compliant way. Larger technology firms such as Microsoft and Apple may exploit data protection to achieve competitive advantage (both firms have declared that they will voluntarily implement the GDPR worldwide). Thus, the ultimate implications of the GDPR on the ad-supported publishing ecosystem may be more nuanced than the negative scenarios being proposed by the online advertising industry. Understanding the impact of the GDPR on online content providers (including, but not limited to, news websites) and their visitors, and specifically websites' ability to provide content (including free content), is the focus of the present paper.

To this end, we developed an infrastructure to allow the collection of data from more than 5,000 websites. Our sample is composed of diverse types of sites, including content providers such as news websites and online magazines, and sites that rely heavily on online advertising. Through this infrastructure, we have been browsing websites repeatedly to collect data on how they interact with the users visiting them, how they manage the collection of users' information (with technologies such as tracking cookies), as well as whether and how they have changed their content offerings. We have also been collecting additional data for these sites from third-party services (such as longitudinal data on their traffic patterns and social media reactions).

The websites included in our sample were selected based on the scope of the regulation. The GDPR applies to every European organization (regardless of country of origin of the organization's customers/users) and provides protection to every European resident (regardless of whether the organization providing the service is based in Europe or not). On one hand, the impact of the regulation might be expected to be more

evident for European websites (which constitute our treatment group) and users; and on the other hand, non-European websites or websites whose user bases are largely non-European (such as U.S. websites i.e. our control group) may be only marginally affected. Thus, both EU-based and US-based websites are included in our sample. To identify the potential impact of the GDPR on downstream economic variables, we collected data for each website at various times, both before and after the GDPR enforcement date. The empirical analysis is therefore based on a difference-in-differences approach: we compare changes in key metrics before and after the enactment of the GDPR in EU vs. US websites. In addition, for certain metrics, we also capture differences between the value of those metrics for EU visitors (that is, visitors recognized by the site as coming from EU IP addresses) vs. US visitors (that is, visitors recognized by the site as coming from US IP addresses).

As noted, we capture changes in the use of tracking cookies and other tracking technologies (e.g. how many third-party cookies are allowed to be set on an individual user's browser); and websites' changes to the use of consent mechanism dialogs (often placed within an overlay on websites to allow users to consent to tracking). While we collect those technical metrics, they do not represent the main focus of our analysis. Rather, our goal is linking those technical changes to their downstream consequences in terms of possible variations in the quantity and quality of websites' content offerings. To measure the quantity of content and quality of content we use proxies previously proposed in the literature, including measures of new material being posted, measures of traffic, and measures of visitors' engagement.

Our objective is to use these metrics to understand possible downstream impacts of the GDPR. Our analysis is ongoing, as data collection is continuing. Our preliminary results suggest that GDPR reduced the number of third-party cookies and tracking responses in both US and EU websites, implying decreased tracking of users by websites in both areas. Furthermore, the enactment of the GDPR may have to some extent neg-

atively affected traffic to EU websites, relative to US sites. However, and importantly, the enactment does not seem to have relatively affected the amount of content that EU websites were able to publish, or the degree of average social media engagement and interaction with such content. These findings aim at contributing to the debate over the economic impact of the GDPR specifically, and the economic impact of privacy regulations, more broadly.

2 Literature Review

This paper builds upon and contributes to three strands of literature: the literature on the economics of privacy (and, in particular, the economic impact of privacy protection and privacy regulation); the economic literature on the online advertising industry; and, more specifically, the small but growing body of economic and non-economic work on the impact of the GDPR.

The economics of privacy literature investigates the trade-offs associated with the revelation or protection of personal information (Acquisti *et al.*, 2016). Within this literature, an important strand of work has focused on the impact of privacy regulation. Policy interventions that regulate the collection or usage of consumer data tend to be aimed at protecting individuals' privacy, but can have a range of nuanced and unpredictable consequences for innovation, market structure, and the economic welfare of different stakeholders. Goldfarb and Tucker (2012) argue that privacy regulation might affect the extent and direction of data-based innovation. However, different works have showed that the impact of privacy regulation can, in fact, be quite heterogeneous, and context specific. In the context of health care, several authors have investigated the impact of privacy regulations on the adoption of electronic medical records (EMR) and the operational effectiveness of health information exchanges (HIE). Miller and Tucker (2009) examine how privacy legislation affects EMR adoption and suggest that

privacy legislation primarily reduces demand for EMRs via the suppression of network effects. On the other hand, Adjerid *et al.* (2015) find that, in the case of HIEs, although privacy regulation can result in a reduction in HIEs' operational effectiveness, if the right privacy incentives are provided to patients, regulation can have a positive impact on the development and adoption of HIEs.

Broadly put, when limits are imposed on the type or amount of data that can be collected and used, industries that are reliant on these data may be affected. The online advertising market offers a clear example: in online targeted advertising, ads are targeted to individuals based on information collected online, usually through the use of tracking technologies. The core idea is that personalized (targeted) ads are more effective than non-targeted ads since consumers see only those ads that are relevant to them, and this provides higher returns for the advertising companies (Evans, 2009). The advertising industry has been quick to complain that restrictions on their ability to collect and use consumer data for targeted advertising is harmful to both advertising companies and consumers (Castro, 2010; IHS Technology, 2015). In a different context, Goldfarb and Tucker (2010) empirically investigated how the earlier 2002 EU 'Privacy and Electronic Communications Directive', which restricted advertisers' ability to collect data on users, affected advertising effectiveness captured by hypothetical purchasing intentions. Their results show that after the regulation, certain types of display advertising were less effective relative to display advertising in other countries. In a related study, Campbell *et al.* (2015) theorized that regulations which require opt-in consent to data collection could affect small and new firms disproportionately and reinforce monopolies. They showed also that users might be more willing to share private information with large companies than with new entrants and small firms.

Within the large body of work on privacy regulation, our paper is related to the recent wave of studies on the impact of the GDPR. Although the regulation was introduced only in May 2018, it has already caught the attention of numerous scholars, in

different fields. Jia *et al.* (2018) focus on the impact of the GDPR on investments in EU emerging technologies; they suggest that, at least in the short-run, the regulation has led to a decrease in such investments for EU companies, compared to US organizations. Similarly, Goldberg *et al.* (2019) examine the effect of the GDPR on European web traffic and e-commerce sales and find that recorded page-views and recorded revenues have fallen by about 10% for EU users. Nevertheless, in a theoretical paper, Lefouili and Toh (2018) conclude that the effect on investments of the GDPR may be mixed. For example, they find that in a fully covered market, although regulating information might reduce investments, it can be socially desirable in a setting where information and quality are not strong complements. Choi *et al.* (2019) investigate consumers' privacy choices with a model in which consumers are required to consent to the collection of their data and consumers are fully aware of the consequences of giving such consent. The authors find that information externalities and coordination failures among users are drivers of excessive loss of privacy.

The impact of the GDPR on websites' technical features is nuanced. Degeling *et al.* (2019) investigate online websites' compliance with the consent requirements imposed by the GDPR. They find that while most websites have adjusted their privacy policies and implemented consent mechanisms, some have not complied and do not provide users with means to consent to tracking. Dabrowski *et al.* (2019) browsed websites from EU and US IP addresses and found that EU-based visitors were less likely to receive persistent cookies compared to US visitors, even as the number of US-based visitors decreased. In the same vein, Urban *et al.* (2020) show that a particular type of cookie - syncing cookies - which allows the exchange of users' information between online advertising actors such as Ad networks and Ad exchanges, decreased across more than 2.6 million websites by approximately 40% around the time the GDPR came into effect. However, they found that the number of syncing cookies slightly increase again over the long-term. In a related study, Sørensen and Kosta (2019) conducted a longitudinal

empirical study to assess the effect of the GDPR on the presence of third parties on EU websites. While they show that the number of third parties did slightly decline after the GDPR, they ultimately conclude that the GDPR may not necessarily be responsible for that effect. Finally, Sanchez-Rola *et al.* (2019) investigate the use of opt-out options by users and find that, despite the presence of the opt-out mechanism, it is still difficult for users to avoid being tracked. Specifically, about 90% of the websites involved in the study placed tracking cookies on users' browsers before they were given the chance to opt-out.

We contribute to these diverse streams of literature by analyzing the impact of regulatory interventions that potentially limit the collection of user data and the tracking of online visitors on variables of critical economic importance for online content providers and, ultimately, their users.

2.1 Theoretical framework

The focus of this manuscript is to investigate the impact of the GDPR on the provision of online content. Ad-sponsored business models have become prevalent among online content websites (Casadesus-Masanell and Zhu, 2013; Goldfarb, 2004; Lambrecht *et al.*, 2014). The more valuable ads are tailored to visitors' preferences which relies rely on the ability to collect personal data. In an advertising context, personal information increases targeting efficiency but at the price of some degree of loss of privacy (Tucker, 2012). Besides allowing more granular targeting, online advertising also has significant cost advantages compared to offline advertising (Goldfarb, 2014). These features make online advertising more efficient and more accountable, in the sense that advertisers can monitor advertising performance and effectiveness through quantitative metrics (Johnson *et al.*, 2017). This creates a self-reinforcing cycle that increases the efficiency of online advertising.

Thus, industries and markets that rely on personal data and profiling might be

impacted by the reduction of the ability to track customers (Goldfarb and Tucker, 2010) imposed by the GDPR with explicit consent mechanisms and associated penalties for non-compliance. Changes in ad effectiveness may in turn affect publishers' revenues, as content providers rely heavily on ad revenue to support their offers of (often free) content to users. And if publishers' profitability is significantly impacted, this could ultimately affect their ability to continue to provide content and sustain its quality. Changes in content quantity and users' engagement (as a proxy of content quality) are the focus of our study.

In a recent empirical study using data from an intermediary for the online travel industry, Aridor *et al.* (2020) find that the total number of consumers observed by the intermediary decreased by 12.5% after the GDPR, suggesting that a significant number of consumers decided to opt-out. The authors additionally find that the remaining set of consumers who decided to not opt-out were more persistently identifiable. Finally, they observed a drop in ad interactions across their data-set, along with an increase by advertisers in the average bids for the remaining observable consumers, leading to a smaller overall decline in revenue.

Research in the online advertising and media literature has investigated the relationship between (changes to) ad-sponsored business models, content providers' incentives, and content provision. Several theoretical studies have argued that when content providers are supported by advertising revenue, they have an incentive to adjust their content to maximize traffic; by so doing, they aim at attracting more advertisers willing to buy ad space on their websites, targeted to specific audiences (Anderson and Gabszewicz, 2006). Empirically, Monic and Feng (2013) investigate changes in the quality of blogs' posts after the implementation of ad-supported business models. They find that the quality of blogs' posts tend to increase because of ad revenue. Shiller *et al.* (2018) investigate whether the increasing adoption of ad blockers by online users might decrease the quality of online content. The authors use traffic at the website level as a

proxy for quality, and find that websites with a high proportion of ad blocking visitors experience a deterioration in traffic ranking relative to websites with fewer ad blocking visitors. Athey *et al.* (2018) show how consumer switching – that is, consumers consuming content from multiple websites – affects advertising strategies and, in turn, increases the competition among publishers, leading to an increase in a publisher's incentives to invest in quality content that attracts a greater share of consumers.

Our contribution to this literature is to study (and compare) the differential effects that the GDPR has on both tracking mechanisms (including differential changes for visitors browsing from EU IP addresses compared to US IP addresses) and on websites' content offerings, for EU websites compared to US websites. To do so, we rely on two different kinds of metrics. First, we collect data to measure how the tracking behavior of websites changes after GDPR comes into effect. Second, we collect different metrics aimed at measuring the ability of websites to produce new content and users' engagement with the websites, that can be thought of as a proxy for quality of the content. These metrics comprise the dependent variables in our analysis. The rationale behind our approach is that the implementation of the GDPR limits the amount of data collected by websites on users. As shown by Johnson *et al.* (2020), this limitation can reduce the quantity of information given to advertisers whose aim is to target ads. In our study, we are able to measure whether the regulation does limit the information flow by analyzing the evolution of tracking cookies and the change (if any) in the amount of information exchanged between websites and advertising networks. As an illustration, a decrease in the amount of available information on users can lead to a decrease in the ability to target highly personalized ads to those users; in turn, this can lead to a change in advertisers' willingness to pay for such ads. If advertisers' willingness to pay changes, we should also observe a change in the revenues earned by content providers from targeted ads.

3 Institutional details

In this section, we provide a brief overview of the GDPR. We focus on the aspects of the regulation that might be expected to directly impact the online advertising ecosystem.

As already discussed, the GDPR could decrease the ability of advertisers to target users with highly personalized ads due to new restrictions placed on data controllers and processors such as ad-tech firms. These restrictions require data processors to justify any actions that involve user data under one of six lawful bases as defined by Article 6 of the GDPR. Two of these, user consent and "legitimate interest" are potentially relevant to advertisers. Of the two bases, obtaining consent may reduce the effectiveness of targeted ads. If a large number of users decide to opt-out (and therefore, do not consent to be tracked), the amount of information available about the users may decrease, making the targeted ads less precise and/or less personalized. Perhaps because of this, some advertisers are claiming that processing user data for targeted advertising is a "legitimate interest" – a term that is only vaguely defined by the text of the regulation (UK Information Commissioner's Office, 2019).

Beyond Article 6, several requirements related to data collection and usage are contained in Article 22 of the GDPR, however it is unclear whether these requirements apply to tracking practices used in advertising. Specifically, Article 22 sets rules for automated decision making by establishing the right of data subjects to not be "subject to a decision based solely on automated processing, including profiling". Such decisions may only occur when they are "based on the data subject's explicit consent". Applied to advertising, it is unclear whether falls under the definition of automated decision making. Some guidance on the application of these rules has been provided by the European Data Protection Board (EDPB) and before it, the Article 29 Working Party (WP29). In an early document released by WP29 on individual decision making and profiling, building profiles for behavioral advertising is presented as an example of an activity governed by the rules of Article 22 (Article 29 Data Protection Working Party,

2017). However, this guidance is not legally binding and it remains unclear whether it will be followed by advertisers.

Under the text and guidelines of the GDPR, we might expect many of the technologies employed in the construction of behavioral profiles to be impacted. While the interpretation of the GDPR is not yet settled, we might expect to see some variations in the implementation of consent mechanisms. Indeed, we observed several occurrences of "tracking walls" within our data set, where a website will restrict access to content until users consent to tracking. This, in part, reflects the differences in the interpretation of the regulation by industry groups such as IAB Europe and advertising firms such as PageFair (IAB Europe, 2017; Ryan, 2017). Until the EDPB or EU Court of Justice provides additional guidance (or enforcement action is taken), it will not be clear which websites are in compliance with the GDPR and which are not. In this study we take advantage of this ambiguity and explore whether differences on the way websites implement the GDPR will have any effect on their economic outcomes.

4 Experimental design

Our experimental design has two components: a sampling strategy for websites whose data we collect; and a set of metrics we collect from each of those websites periodically.

We detail in subsection 4.1 our website sampling methodology and in subsection 4.2 the collected metrics.

4.1 Websites Sample Selection

We constructed a large longitudinal panel of websites focusing on websites located in the EU and in the US. The panel includes websites of different types, including content providers such as news websites and online magazines, but also shopping websites. We visited all sites in our sample periodically and repeatedly using a web privacy measure-

ment framework (OpenWPM). We refer to each of these data collection instances as a "wave". In each of the waves, we attempt to browse all the websites included in the sample and collect the same set of metrics. In addition to data collected by visiting the websites themselves, during each wave we also collect from their parties data associated with each of the websites - such as data on their traffic patterns and social media reactions. In this subsection, we describe how we constructed the panel.

As our goal was to uncover the impact of the GDPR on both major and minor content providers in European countries and in the US, we included in our sample both top ranked and long-tail websites across different content categories.

We started our sampling process by using 2018 Alexa data to choose top ranked websites. Alexa data provides the top 500 websites from various geographical areas (including Global, Germany, France, Italy, Spain, Netherlands, and USA.)[3] and five content categories (News, Sports, Society, Health, and Games).[4]

Alexa's top 500 websites by country correspond to the websites visited most by users in a country, as opposed to the most popular websites that are based in that country. To include the top websites based in each of our countries of interest, we used Alexa's global top 1 million websites to complement the dataset with the top 500 websites in various top-level websites, such as *.de, .fr, .it, .es,* and *.nl*.

After examining the global ranking of the websites included in the sample produced using the criteria outlined above, we noticed that the sample was heavily concentrated in highly ranked websites and that websites ranked below 200,000 were not included. Thus, we decided to include an additional random sample of websites ranked between 200,000 and 1 million. To do so, we included 500 random websites for each 100k websites ranking interval, i.e. 500 websites ranked between 200k and 300k, 500 websites ranked between 300k and 400k, and so on until reaching 1 million. To avoid

[3]See, below, how the "location" of a website is defined.

[4]Our sample includes also websites from the UK and from Australia, as controls. As our data analysis is ongoing, we do not yet discuss results for those control websites in this version of the manuscript.

websites that were not relevant for our analysis, we considered only websites in the following top-level websites: *.au, .de, .fr, .uk, .it, .es, .nl, .com, .net* and *us*. The resulting sample included 11,254 websites. We further cleaned this sample so to eliminate websites that only get a minor fraction of their visitors from EU countries or the US, despite the fact that they were among the most popular websites in one of our countries of interest. For example, the Russian shopping website *avito.ru* was the 52nd most visited website in The Netherlands in May 2019. However, visitors from The Netherlands account for less than 2% of all avito.ru's visitors, while visitors from Russia account for roughly 85%. Therefore, although the website is popular in at least one EU country, it would be unreasonable to assume it will significantly change its behavior due to a European regulation considering that it is a Russian website that gets most of its visitors from Russia. Finally, we also scrutinized the content categories of the remaining websites. We noticed that the content categories provided by Alexa were often inconsistent. Thus, we obtained categories information using SimilarWeb,[5] which in our experience provided a more robust categorization. We excluded from the sample all websites categorized as providing adult content, or not assigned to any category.

The resulting sample contains 5,474 websites. Table 1 presents the distribution of websites by global ranking, while Table 2 shows the distribution by content (sub)category (as reported by SimilarWeb). In both tables, the country of a website was determined by the location of its headquarters as reported by SimilarWeb. When this information was not available, we defined the country of a website by using the website's top-level domain country of origin. In the case where neither the country nor the top level domain are available, we assigned the country to where most visitors originated from. In Table 1, a large number of websites are concentrated in the top 99,999 ranking where EU websites represent the majority in this category. Table 2 shows that a large number of websites belongs to the *News and Media category* disregarding the

[5]See `https://www.similarweb.com/`.

country.

Table 1: Number of websites included in sample by global ranking

Ranking	Nr. websites	EU websites	US websites
0 – 99,999	3920	2477	1443
100,000 – 199,999	405	190	215
200,000 – 299,999	252	103	149
300,000 – 399,999	183	79	104
400,000 – 499,999	178	77	101
500,000 – 599,999	124	50	74
600,000 – 699,999	117	40	77
700,000 – 799,999	119	23	96
800,000 – 899,999	122	58	64
Over 900,000	54	16	38
Total	5474	3113	2361

Table 2: Number of websites included in sample by content sub-categories

	Total	EU websites	US websites
News and Media	946	487	459
Arts and Entertainment	528	327	201
Shopping	514	416	98
Business and Industry	413	260	153
Health	326	87	239
Internet and Telecom	291	189	102
Games	325	106	219
Sports	307	137	170
Career and Education	310	214	96
People and Society	236	103	133
Finance	189	145	44
Travel	177	128	49
Computer and Electronics	174	92	82
Law and Government	141	75	66
Autos and Vehicles	146	102	44
Food and Drink	109	64	45
Gambling	79	53	26
Reference	77	48	29
Beauty and Fitness	42	11	31
Science	42	15	27
Pets and Animals	32	11	21
Books and Literature	29	19	10
Home and Garden	20	13	7
Recreation and Hobbies	21	11	10
Total	5474	3113	2361

4.2 Data Collection

For each website in our sample, we collected an array of metrics over a period of 19 months (from April 2018 to October 2019), that includes a total of twelve waves.

We focus on two sets of metrics. The first corresponds to data we directly mine from the websites in our samples and that capture how users are tracked. The goal of collecting these metrics is to determine how websites change their tracking behavior after the implementation of the GDPR. We refer to these variables as "technical variables" (see Section 4.2.1). These technical variables were captured by browsing each website from two different IP addresses, one located in Europe (France) and one in the US.[6] This experimental design allows us to compare, before and after the enactment of the GDPR, whether and how websites adapted their data collection behavior according to the geographical location of a visitor.

The second set of metrics is obtained from third parties repositories and are aimed at measuring the quantity of content offered by all the websites in our sample and users engagement with such content, as a proxy for its quality. We refer to these variables as "content variables" (see Section 4.2.2). In this case, the data does not change as function of the country of the visitor. However, we do expect to find differences depending on the location of registration of the website, as websites registered in different locations (EU vs US) should be affected differently by the GDPR.

Both sets of metrics are crucial for our analysis. Although our investigation ultimately focuses on whether the implementation of the GDPR had downstream effects on websites' ability to sustain quantity and quality of content, understanding the intermediate impact that the GDPR may have had on websites' ability to track and manage users' data is instrumental to our primary analysis. In other words, while content variables are the dependent variables of our analysis, technical variables are useful

[6]We also collected data with IP address from two other countries i.e. UK and Australia, as robustness checks. As the analysis is ongoing, we do not yet present results for these IP addresses in this version of the manuscript.

controls.

4.2.1 Privacy & Ads variables

To simulate user browsing and investigate how the websites interact with the user, we rely on OpenWPM, a web privacy measurement framework (Englehardt and Narayanan, 2016). This framework is implemented within an instrumented web browser that automates the process of visiting a set of websites and records a series of variables related to websites' handling of personal information during those visits. Using this framework, we visited each website in our sample periodically for a period of 19 months. We call each round of visits to all websites a "wave of data collection". During each wave we visited each website twice, once with an US IP address and once with an EU IP address. The goal was to contrast how the same website may handle private information differently based on the user location.

Through OpenWPM, we collected different metrics that measure the data collection practices of the websites in our sample. We further processed some of this data to construct additional metrics related to advertising shown on those websites. Using scripts included in popular ad blockers, we flagged advertising content within the HTML content we extracted via OpenWPM. An ad blocker is a small piece of software or module incorporated into a user's browser (Add-on) that prevents the display of banners and other advertising formats. Ad-blockers filter advertisements by recognizing the advertising tags of the main ad servers and advertising networks. We cross-referenced the data we collected from OpenWPM with these filtering lists (blocklists).[7]

In particular, We rely on two blocklists; Adblock Plus[8] and Disconnect.[9] which establish identification and classification rules for advertising and tracking entities.

In the rest of this section, we describe in detail the various technical variables we

[7]List inside ad-blocker add-on to block unwanted content like advertising.

[8]`https://adblockplus.org/fr/subscriptions`. Last retrieved, February 2020

[9]Disconnect is a free extension for the web browser responsible for blocking trackers from web pages that the user visits. `https://disconnect.me/`

collected.

Cookies: Cookies are small files stored on visitors' browsers and often embedded on websites to provide additional functionality. There are two main types of cookies: 1st party and 3rd party cookies. The variable *1st party cookies* measures the cookies which are set by the website being browsed. The variable *3rd party cookies* represents cookies that are set by entities other than the original website, and that could be used to track users' behavior across different websites in order to construct users' profiles aimed, in part, at improving behaviorally targeted ads. The variable *Tracking Cookies* counts the number of tracking cookies (from *Adblock Plus Easylist*) while *Session cookies* includes temporarily saved information whose content is deleted after the browsing session is completed or the web browser is closed. Finally, *Persistent cookies* are cookies which remain available on the browser for more than 30 days.

HTTP responses: To get a better overview of third-party cookies, we also collected HTTP responses. HTTP responses measure all the information exchanged between the browser and the websites that are visited by a user. By capturing all HTTP responses, we were able to identify the number of responses linked to advertising and the subsequent advertising networks which serve a website. The variables *3rd party resp*, *Tracking resp* are continuous and correspond respectively to the number of responses made by the browser to websites other than the website being browsed and the number of responses made by all tracking entities on a website. We also used HTTP responses to identify websites with consent mechanisms. The consent mechanisms are the means through which users should be able to express their choice regarding whether they wish to opt-in or opt-out of data collection. We record the presence of a *Consent mechanism* on a website using a dummy variable. We were able to identify website con-

sent mechanisms by matching HTTP responses with a crowd-sourced list of elements,[10] equivalent to the Adblock blocklist, but for consent mechanisms following Eijk *et al.* (2019) methodology.

Advertising content length: We were also interested in measuring the length (in bytes) of certain types of websites' html content which we collected while browsing our sample of websites. In particular, we were interested in the length of advertising content. The main purpose of collecting this metric is to get information about the amount of ad content available on a given website's homepage. In more detail, the content length variable measures the size of the response we received from the websites and includes any advertising or image content contained in the website. To specifically capture advertising content quantity (which we identified via the *Adblock Easylist*), we used the variable *Ads length (KB)*. This variable gives information on the size in kilobytes of the quantity of advertising content on a website.

4.2.2 Content variables

Analyzing the impact of the GDPR on economic outcomes from a provider's perspective, requires us to go further in the data collection by getting additional information related to downstream economic outcomes: content quantity and users' engagement with such content.

We measure users' engagement using websites' traffic metrics and social media reactions, as explained below. The underlying premise is that if the quality of the content provided by the website decreases, users may try to substitute for other content and, therefore, we should observe a decrease in the number of visits to a given website. This is similar to the strategy used by Shiller *et al.* (2018) to measure websites' quality. Specifically, we use *Reach per million* as a measure of the number of users visiting a website. This variable estimates the average number of users visiting a website per

[10] www.i-dont-care-about-cookies.eu/

million random users on the internet. In addition, we also use *Page views per user*, which represents the number of pages viewed per user on a website. In addition, we augment the traffic variables with data on the number of social media "reactions" related to content published on the websites in our sample. We use data from Facebook Graph API in line with Cagé *et al.* (2015) methodology, that used the same metric as a proxy of quality for online news websites. Specifically, we collect the number of reactions on the Facebook platform for each new url of content posted by the websites in our sample (as retrieved via GDELT - see further below), and calculate the average number of Facebook reactions for each new url of content across all the new urls of posted content on a given website during a given wave. Such reactions can be used to measure users' engagement with a piece of content, and a proxy for content quality.

To measure content quantity, we capture the total number of new urls of content published by each website in our sample in the week surrounding each observation from OpenWPM. Because we visit each website multiple times to construct our longitudinal data set, we collect multiple observations of the new urls counts for each website over time. For this purpose, we rely on the *Global Database of Events, Language, and Tone* (GDELT) to estimate the quantity of content for each website. GDELT provides links to content pages going back to 2015 from both domestic (US) and international sources. This database allows us to collect the total number of urls on a given website three days before and after each observation collected with OpenWPM. We rely on GDELT because it allows us to retrieve and date content data retrospectively and because it overlaps with a large number of websites in our sample.[11]

[11] `gdeltproject.org`

Table 3: Descriptive Overall

	Mean	Std. Dev.	Min	Max	N
		Technical variables			
Cookies:					
1st party Cookies	8.93	7.03	0.0	61.0	122843
3rd party Cookies	21.26	33.56	0.0	351.0	122843
Tracking Cookies	6.82	10.75	0.0	126.0	122843
Persistent Cookies	23.75	30.14	0.0	314.0	122843
Session Cookies	4.03	4.31	0.0	44.0	122843
Response:					
3rd parties Resp.	138.89	200.87	0.0	5186.0	113463
Tracking Resp.	20.37	29.80	0.0	1157.0	113463
Advertising:					
Ads length (KB)	248.82	683.38	0.0	111046.5	122504
Consent Mechanism:					
Consent Mechanism	0.26	-	0.0	1.0	122843
		Content variables			
Content Quantity:					
GDELT URLs	245.27	585.15	1.0	14171.0	15787
Content quality:					
Reach per million	434.56	9686.57	0.1	607471.4	58915
Page views per User	2.90	2.26	1.0	107.9	58915
User engagement:					
FB Average Reaction	473.03	2841.37	0.0	263908.0	15422

Notes: This table presents the descriptive statistics for the overall sample.

5 Descriptive evidence

As argued in Section 2, the GDPR is likely to reduce the ability of websites (as well as of other players in the online advertising ecosystem) to track and collect users' data, creating a cascade of effects that may lead to a decrease in online advertising revenues for various stakeholders (including online content providers) and, ultimately, to a reduction in the quantity and quality of online content.

To investigate these cascading effects, in this section we provide a series of descriptive statistics that capture whether 1) the websites in our sample changed the way they handle private information after the GPDR became effective, and 2) whether, in turn, those changes ended up affecting websites' content quantity and user engagement. As noted in Section 4 we use the term "technical variables" to refer to metrics that inform

us about changes in the way websites handle private and advertising data, and we use the term "content variables" to refer to variables we use as proxy for changes in the quantity of content provided and users' engagement with such content. In Section 5.1, we first present descriptive statistics showing how websites located in the EU or the US (denoted in the tables and figures *EU websites* and *US websites*) handled data on visitors from an EU location or a US location (denoted *EU IP* and *US IP*), before and after the implementation of the GDPR. Then, in Section 5.2, we present descriptive statistics describing changes in content quantity and users' engagement with content posted by EU and US websites before and after implementation. Finally, in Section 6, we use a difference-in-differences estimation model to estimate the impact of GDPR on content variables, controlling for various technical variables such as cookies, advertising length, and consent mechanisms.

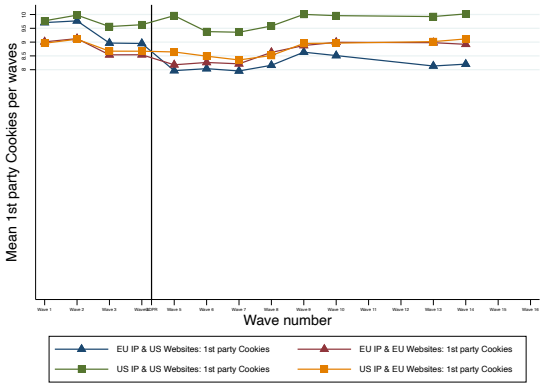
5.1 Tracking and & Advertising

To explore the effect of the GDPR on how websites handle private and advertising related data, we first analyze the number of (1st party, 3rd party, tracking and advertising) cookies and HTTP responses presented by the website during browsing.

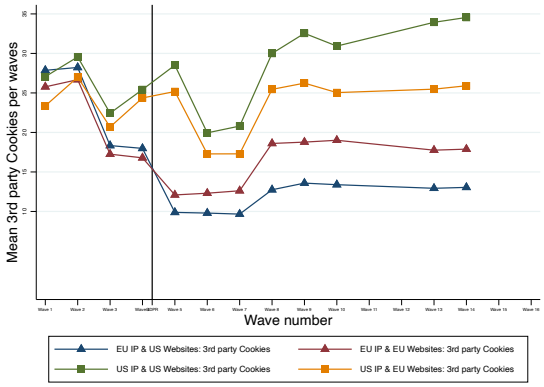
Figure 1 displays the average number of first-party cookies (i.e. those issued by the website itself) and the average number of third-party cookies (i.e. those issued by parties other than the website). Generally, the average number of third-party cookies is higher than first-party cookies. More importantly, while the number of first-party cookies remains largely unchanged regardless of the visitors' IP address, we observe a large impact of the GDPR on third-party cookies.

Fig. 1 *1st and 3rd parties Cookies*

a *First party cookies*

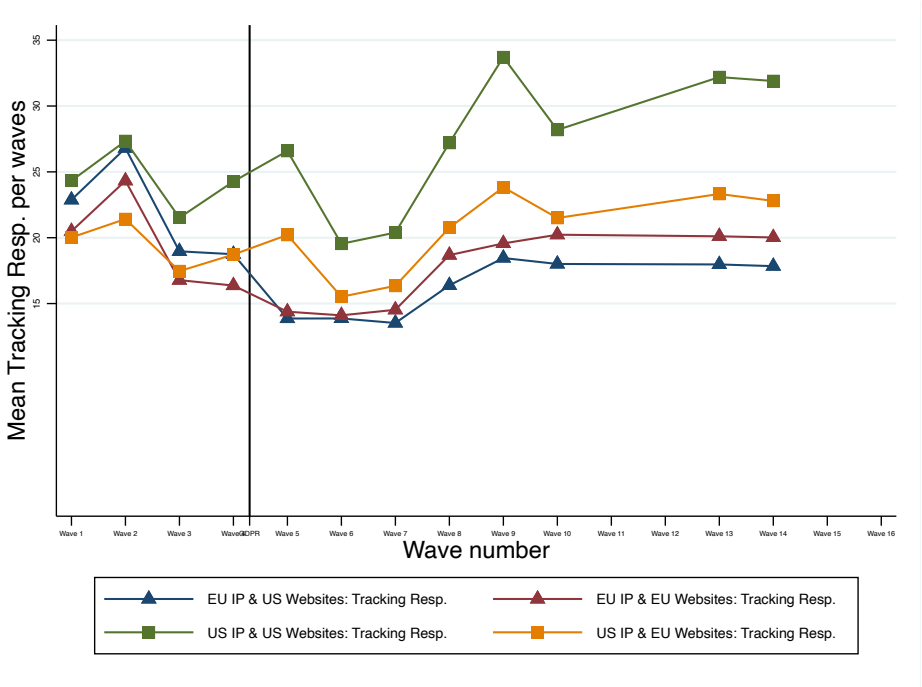


b *Third parties cookies*



In particular, Figure 1b shows differences in third-party cookies between EU and US websites according to the IP location of the visitor. For EU IP addresses, use of third-party cookies is greatly reduced after the GDPR became effective. However, the regulation does not seem to have affected the number of third-party cookies in the case of US IP addresses, independently of the website localization. This pattern is plausible, considering the nature of the GDPR. First-party cookies are less likely to be related to tracking and advertising and thus their use should not be affected by the regulation. Instead, third-party cookies are probably related to tracking and advertising and it is therefore reasonable to expect a reduction in the case of EU IP addressees but not necessarily in the case of US IP addresses.

Fig. 2 *Tracking response*



The results in Figure 2 are more striking. The figure depicts the average number of tracking responses that we identified by classifying responses using the Adblock Plus and Easylist lists of known trackers. The figure shows that the introduction of the regulation greatly reduced the number of tracking responses for visitors originating from the EU (EU IP) on both EU and US websites. However, in the case of visitors originating from the US (US IP) the number of tracking responses remained fairly constant or even increased, especially when visiting US websites. This is further evidence that websites seem to be adhering to the regulation for EU visitors, and that EU websites are being more cautious about their tracking behavior compared to US websites.

Figure 3 depicts the average ads length in Kilobytes on websites, by website location and IP address of the visitor. The implementation of the regulation has had an evident effect on Ads length for EU visitors to US websites; there has been a net decline since GDPR became effective. In contrast, there have been no evident changes in the case of US visitors to US websites. We also note a spike around wave 9 for which we do not yet have a firm explanation. We believe it may be related to an increase in

advertising due to the end of year holiday's shopping season (wave 9 was collected a few days before Christmas).

Fig. 3 *Ads length (KB)*

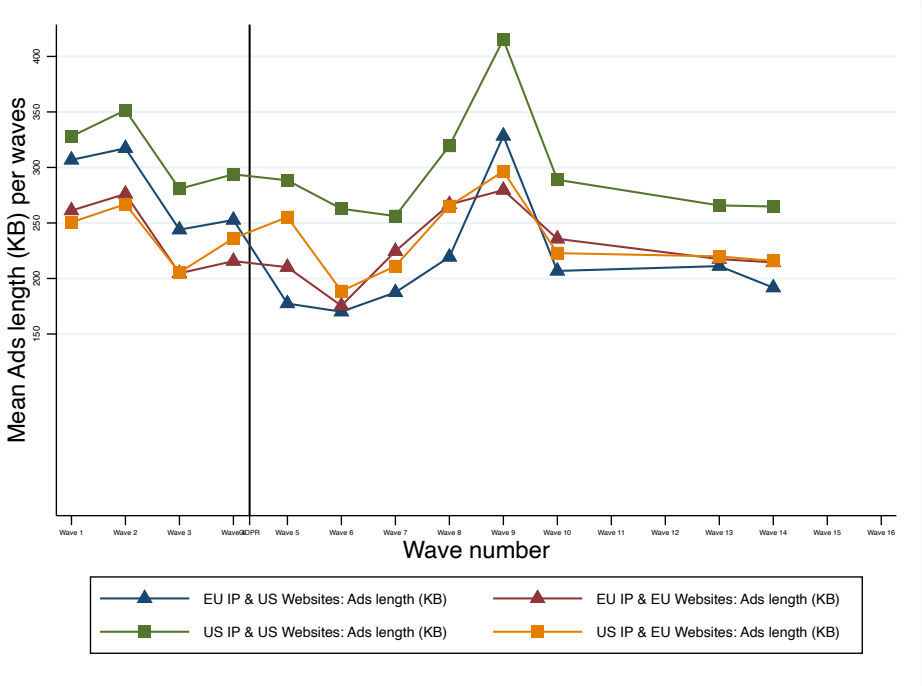
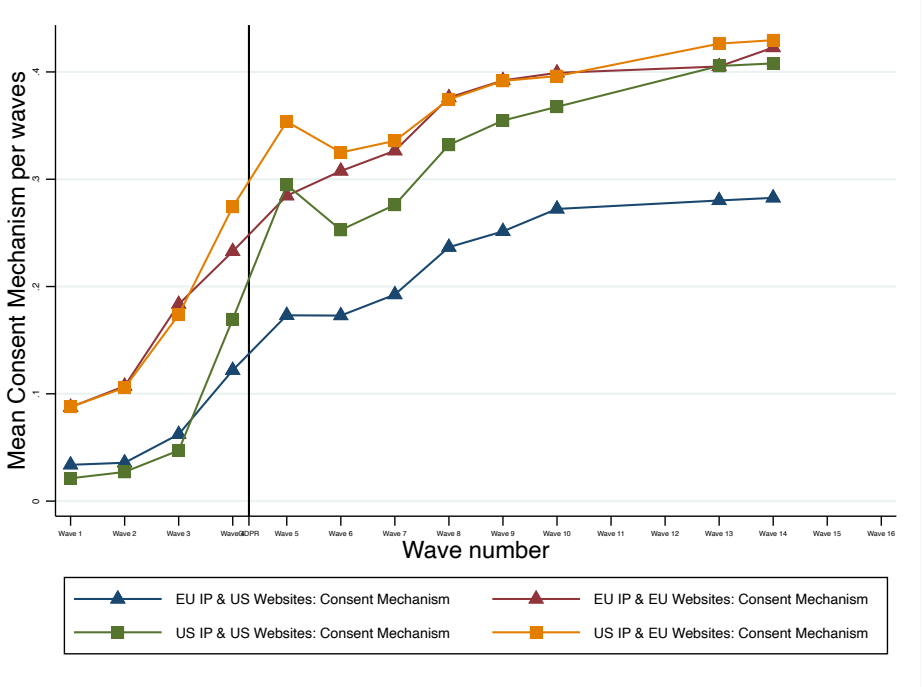


Figure 4 describes the percentage of websites in our sample that provided a consent mechanism to their users. EU websites seem to have been better prepared for the introduction of the GDPR; a higher percentage of EU websites had included a consent mechanism before the law was implemented. However, US websites quickly caught up - although, somewhat surprisingly, we observe a smaller fraction of websites offering a consent mechanism for EU visitors on US websites (compared to US visitors on US websites). This might be because, rather than providing a consent mechanism in order to be compliant with the GDPR, some US websites chose to block EU visitors or to not track them at all (Something that we are, in fact, currently exploring in our ongoing analysis.).

Fig. 4 *Pct of website with consent mechanism*



Tables 4 and 5 present the results of t-tests comparing the means of our privacy-related variables before and after the GDPR, for both US and EU IPs. Table 4 considers US websites and Table 5 focuses on EU websites. In both cases, the effect of the GDPR is much stronger for EU IP addresses than for US IP addresses, suggesting that the enactment of the GDPR was associated with significant changes in tracking for EU visitors with spillover effects on US visitors. While there is a reduction in the number of third-party cookies related to EU IP addresses on both EU and US websites, the opposite effect is observed for US visitors browsing US websites – i.e., there is an increase in the number of third-party cookies. The case of *Tracking resp.* requests is even more telling: the number of requests decreased for visits originating in the EU (EU IP) and increased for visits originating in the US (US IP) on both EU and US websites.

Table 4: T-Test before and after the GDPR for US websites

	US IP			EU IP		
	Bfr GDPR	Aft GDPR	Diff.	Bfr GDPR	Aft GDPR	Diff.
1st party Cookies	9.70	9.79	0.09	9.36	8.20	-1.16***
3rd party Cookies	25.70	29.17	3.47***	23.27	11.87	-11.40***
Persistent Cookies	27.89	31.00	3.11***	25.99	15.26	-10.74***
3rd parties Resp.	157.42	183.90	26.48***	142.77	103.05	-39.71***
Tracking Resp.	23.38	27.71	4.33***	21.18	16.23	-4.95***
Ads length (KB)	310.03	298.07	-11.97	281.16	210.81	-70.35***
Consent Mechanism	0.05	0.34	0.29***	0.06	0.23	0.17***

Notes: This table reports the averages of the variables and highlights the differences before and after the GDPR for US websites. Column 1 to Column 3 present the results obtained by browsing the websites with a US IP address. Column 4 to Column 6 present the results obtained by browsing with an European IP. $*p < .1$, $**p < .05$, $***p < .01$

Table 5: T-Test before and after the GDPR for EU websites

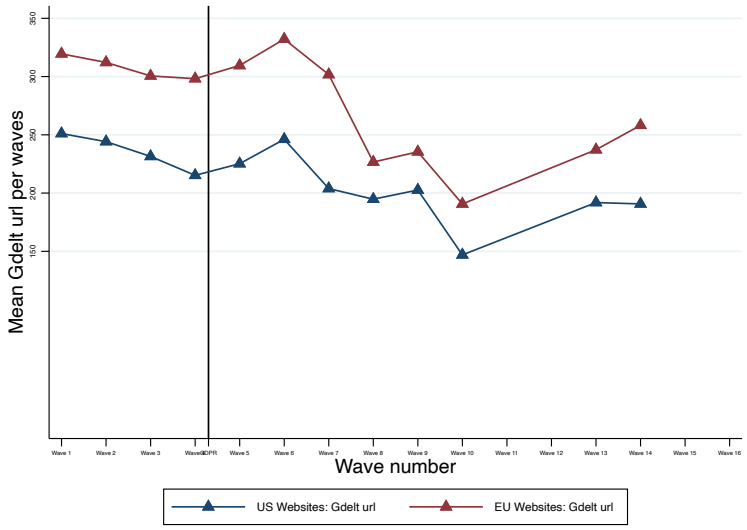
	US IP			EU IP		
	Bfr GDPR	Aft GDPR	Diff.	Bfr GDPR	Aft GDPR	Diff.
1st party Cookies	8.85	8.76	-0.09	8.81	8.63	-0.18**
3rd party Cookies	23.28	23.78	0.50	21.75	16.11	-5.63***
Persistent Cookies	25.32	25.79	0.48	24.18	19.19	-4.99***
3rd parties Resp.	131.57	146.20	14.63***	129.38	124.09	-5.28**
Tracking Resp.	18.79	20.66	1.87***	18.78	17.69	-1.09***
Ads length (KB)	235.46	236.69	1.24	240.16	227.56	-12.60*
Consent Mechanism	0.15	0.38	0.23***	0.15	0.36	0.21***

Notes: This table reports the averages of the variables and highlights the differences before and after the GDPR for US websites. Column 1 to Column 3 present the results obtained by browsing the websites with a US IP address. Column 4 to Column 6 present the results obtained by browsing with an European IP. $*p < .1$, $**p < .05$, $***p < .01$

5.2 Content Quantity and Quality

As noted, we measure the impact of the GDPR on content quantity using the GDELT database. Specifically, we measure quantity as the number of new urls of content posted by the websites within our sample (*GDELT URLs*). Figure 5 shows that there was an initial decline in the number of new URLs published by both EU and US websites immediately after the enactment of the GDPR; nevertheless, the numbers seem to slowly recover after a few months.

Fig. 5 *New urls (GDELT URLs)*



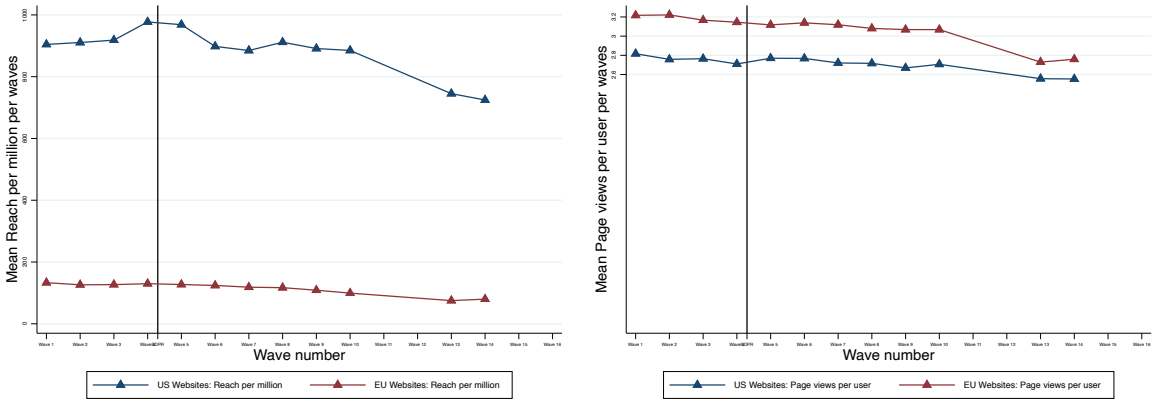
We measure the impact of the GDPR on websites' traffic and social media engagement as a proxy for content quality. Specifically, we use websites' *Reach per million* and *Page views per user*. In addition, we use the number of social media "reactions" related to new content published on the websites in our sample.

In the case of reach (figure 6a), defined as the fraction of Alexa users who visited the website on a per million basis, we observe that EU websites have remained fairly stable, but US websites apparently experienced a decline.

Figure 6b depicts the number of page views per user and suggests a different dynamic. While the pattern may appear fairly stable for US websites, there appears to be a small downward trend for EU websites. Indeed, our statistical analysis, presented below, reveals that EU websites experienced a slight decline in the number of page views. A possible explanation for this is that the implementation of consent mechanisms had a negative effect on users' engagement, because consent pop-ups can be obtrusive and time consuming.

Fig. 6 *Websites Traffic*

a *Reach per million* **b** *Page views*



Finally, for all *GDELT URLs*, we collected the number of reactions on Facebook as a proxy for quality within this subsample (*FB Average Reactions*). Figure 7 shows that the number of reactions on Facebook initially remain stable after the GDPR – although there are spikes (particular for US websites) in later waves.

Fig. 7 *FB Average Reactions*

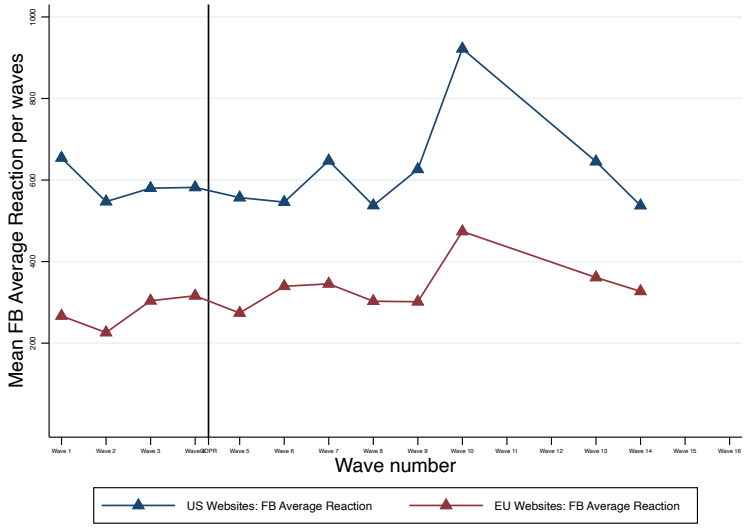


Table 6 compares Alexa's *reach per millions, page views per user* and new urls (*GDELT URLs*) for EU and US websites before and after implementation of the GDPR

using t-tests. In the case of EU websites, almost all the variables are negative and significant except for the *FB Average Reactions*, that is positive. However, and more interestingly, we observe a similar effect for US websites based on GDELT data and *Page views per user*. This suggests that content production may have also been negatively affected across US websites following the implementation of the GDPR.

Table 6: T-Test before and after the GDPR for EU and US websites

	US websites			EU websites		
	Bfr GDPR	Aft the GDPR	Diff.	Bfr GDPR	Aft GDPR	Diff.
GDELT URLs	236.43	200.86	-35.57***	308.14	261.76	-46.38***
Reach per million	926.62	860.67	-65.95	129.03	105.66	-23.37***
Page views per user	2.76	2.68	-0.08***	3.19	3.00	-0.18***
FB Average Reactions	592.19	676.50	84.31	276.71	340.30	63.59*

Notes: Column 1 to Column 3 presents the results for EU websites. Column 4 to Column 6 presents the results for US websites

6 Estimation of the GDPR impact on website traffic and social media engagement

The descriptive evidence we have presented so far suggests not only significant changes in websites' handling of visitors' data following the GDPR, but nuanced and complex variations in websites' content quality and quantity. We use a difference-in-differences (DID) model to tease out potential differences in content quality and quantity after the GDPR, for US and EU websites in a framework that controls for website specific features and time-specific characteristics. The specification of our regressions is as follows:

$$Y_{i,t} = \beta_0 + \beta_1 Post\ GDPR \times EU\ Websites_{i,t} + \delta X_{i,t} + \omega_t + \mu_i + \epsilon_i \qquad (1)$$

In equation (1), $Y_{i,t}$ represents our variable of interest for a website i at wave t; $X_{i,t}$ corresponds to a vector of privacy-related control variables; ω_t is a vector of time fixed effects, and μ_i is a vector of website fixed effects. $Post\ GDPR \times EU\ Websites_{i,t}$ is equal to 1 if the website i is a EU website and wave t was collected after the GDPR became effective, and 0 otherwise. Standard errors ϵ_i are clustered at the website level. In this framework, the coefficient β_1 corresponds to the DID estimator of the effect of the implementation of the GDPR for websites based in the EU. We test several specifications, in order to capture the heterogeneous effect of the GDPR on online content providers. Among them, we focus on News and Media websites (which, due to their reliance on advertising to monetize content, should be more likely to be affected by the GDPR), as well as on websites that, based on our own data, were more likely to rely on advertising before the GDPR.

Table 7 shows the difference-in-differences estimation of the effect of the GDPR

on content quantity and quality, for EU websites relative to US websites. Column (1) presents the results for the metrics used as proxy for content quantity as dependent variables, namely: *Number of GDELT URLs*. Columns (2) to (4) present the results for the metrics used as proxy for content quality as dependent variables, namely: *Log User reach per million*, *Log page views per user* and *FB Average Reactions*. The set of control variables include: *Ads length - EU IP*, *3rd party Cookies - EU IP-*, *Persistent Cookies - EU IP-*, and *Consent Mechanism - EU IP-*, which were collected from an EU IP.[12] Our control variables, along with websites fixed effect, allow us to measure how the variation in handling advertising and tracking by websites affect our dependent variables. In other words, we can control for the cascade effect of GDPR on technical variables on our proxies of content.

In the specifications Column (2) and Column (3), our variable of interest *EU Websites* $\times$ *Post GDPR* is negative and significant. The results presented in Column (2) suggest that after the GDPR, EU websites' reach was negatively affected; in other words, EU websites are getting fewer visitors compared to US websites. Additionally, Column (3) suggests that visitors of EU websites are browsing fewer pages per visit, after the GDPR. One possible interpretation is that the reduction in the number of pages visited on a given website may be a signal of reduction in the quality of the content offering. If the quality of the content is reduced, users may decide to spend less time on the website and divert their attention to other websites. Another plausible explanation is that when users visit EU websites, they now encounter consent notices or requests, and even consent and tracking walls, in some cases. This may lead viewers to leave the page instead of expressing their consent choices. We are currently investigating these possible explanations.

Importantly, however, we do not find any effect in terms of GDELT URLs (1) (one

[12]The estimation of the content is at the website level and not at the website IP level. Therefore, we choose to control for technical variables collected using an EU IP address to have a better idea of the GDPR effect.

of our proxies for content quantity), nor Facebook reactions (4) (one of our proxies for content quality). In short, the results suggest that the enactment of the GDPR may have, to some extent, negatively affected traffic to EU websites, relative to US sites, but has not relatively affected the amount of content that EU websites are able to publish, or the degree of average social media engagement and interaction with such content. Our expectation of these variables is that they are strongly correlated with content quality, considering that the goal of these sites is to attract and retain viewers.

Table 7: Diff-in-diff regressions - Content dependent variables

	Content Quantity	Content Quality		
	(1) Log Number of GDELT URLs	(2) Log User Reach per million	(3) Log Page view per User	(4) FB Average Reaction
EU Domains × Post GDPR	0.041	-0.040***	-0.029***	-66.235
	(0.032)	(0.011)	(0.007)	(51.343)
Ads length (KB) -EU IP-	-0.000	-0.000	0.000*	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
3rd party Cookies -EU IP-	-0.003	-0.003***	-0.001*	-2.882
	(0.002)	(0.001)	(0.000)	(3.065)
Persistent Cookies -EU IP-	0.003	0.004***	0.001	4.140
	(0.002)	(0.001)	(0.001)	(3.610)
Consent Mechanism -EU IP-	-0.057**	-0.024**	0.013**	29.252
	(0.026)	(0.010)	(0.005)	(36.066)
Constant	4.058***	3.093***	0.976***	312.461***
	(0.022)	(0.010)	(0.005)	(45.068)
Waves Fixed effect	Yes	Yes	Yes	Yes
Website fixed effect	Yes	Yes	Yes	Yes
Std. err Websites level	cluster	cluster	cluster	cluster
Obs	15731	58899	58899	15084

Notes: Estimates from the DID estimation for all websites in our sample. Column (1) presents estimation for the *Log Number of GDELT URL* dependent variable. Column (2) reports estimation with *Log User Reach per million* as dependent variable. Column (3) presents *Log Page Views per User* as dependent variable. Column (4) reports estimation with *FB average Reaction* as dependent variable. All estimations include waves and website fixed effect.
Standard errors in parentheses and clustered at the website level.
Significance levels: $*p < .10$, $**p < .05$, $***p < .01$.

6.1 News and media

A category of websites likely to be particularly affected by the implementation of the GDPR are News and Media websites, as they tend to offer free content in exchange for users' attention and users' data. Table 8 presents the difference-in-differences estima-

tions of the effect of the GDPR on content quantity and quality for the sub-sample of News and Media websites. The coefficient of *EU Websites* × *Post GDPR* is negative and significant in column (3). The effect is stronger for *Log Page Views per user* than in table 7, suggesting that the decrease in the number of page views per user for EU websites is stronger for news websites. However, the effect on Reach is not significant for News and Media websites – indicating that the GDPR has not reduced the number of visitors on EU News and Media websites. This result might suggest that visitors don't have an alternative way to consume European News and media, so the reach stays unchanged. The number of page views per visitor decreased, which might indicate that users are spending less time in websites either because of a decrease in quality, or because visitors abandon websites after experiencing requests for expressing consent choices. On the other hand, the results also suggest a possible *increase* in quantity of content for EU news websites relative to US websites - although the results are only significant at the 10 per cent level.

Table 8: Diff-in-diff regressions: For content variables dependent on the News and Media subsample

	Content Quantity	Content Quality		
	(1) Log Number of GDELT URLs	(2) Log User Reach per million	(3) Log Page view per User	(4) FB Average Reaction
EU Domains × Post GDPR	0.078*	-0.012	-0.040***	7.275
	(0.042)	(0.020)	(0.012)	(66.184)
Ads length (KB) -EU IP-	-0.000	-0.000**	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
3rd party Cookies -EU IP-	-0.002	-0.000	-0.000	-2.588
	(0.002)	(0.001)	(0.001)	(2.401)
Persistent Cookies -EU IP-	0.001	-0.000	0.000	3.952
	(0.002)	(0.001)	(0.001)	(2.684)
Consent Mechanism -EU IP-	-0.074**	-0.008	0.008	50.948
	(0.035)	(0.017)	(0.008)	(43.986)
Constant	5.209***	3.796***	0.673***	318.917***
	(0.028)	(0.016)	(0.009)	(59.148)
Waves Fixed effect	Yes	Yes	Yes	Yes
Website fixed effect	Yes	Yes	Yes	Yes
Std. err Websites level	cluster	cluster	cluster	cluster
Obs	8321	10303	10303	8148

Notes: Estimates from the DID estimation for news and media websites Column (1) presents estimation for the *Log Number of GDELT URL* dependent variable. Column (2) reports estimation with *Log User Reach per million* as dependent variable. Column (3) presents *Log Page Views per User* as dependent variable. Column (4) reports estimation with *FB average Reaction* as dependent variable. All estimations include waves and website fixed effect.
Standard errors in parentheses and clustered at the website level.
Significance levels: $*p < .10$, $**p < .05$, $***p < .01$.

6.2 Websites that rely on advertising

We ran the same analysis for the sub-sample of websites that relied on advertising to monetize their content before the GDPR. The sub-sample includes all the websites that, at any time prior to the GDPR, have an advertisement content length greater than zero. Table 9 shows statistically significant reductions in the user reach per million and in the number of page views per user for EU sites. This suggests a decrease in traffic also for this type of EU websites, relative to US sites, after the implementation of GDPR. However, we find no statistically significant change in the quantity of new content published or in Facebook reactions.

Table 9: Diff-in-diff regressions: For content variables dependent on the subsample of websites relying on advertising

	Content Quantity	Content Quality		
	(1) Log Number of GDELT URLs	(2) Log User Reach per million	(3) Log Page view per User	(4) FB Average Reaction
EU Domains × Post GDPR	0.053	-0.010	-0.037***	-69.765
	(0.038)	(0.016)	(0.009)	(53.242)
3rd party Cookies -EU IP-	-0.001	-0.000	-0.001	-1.884
	(0.002)	(0.001)	(0.000)	(3.136)
Persistent Cookies -EU IP-	0.001	0.000	0.000	2.978
	(0.002)	(0.001)	(0.001)	(3.702)
Consent Mechanism -EU IP-	-0.065**	0.015	0.013**	26.761
	(0.028)	(0.012)	(0.006)	(33.999)
Constant	4.569***	3.629***	0.791***	327.445***
	(0.027)	(0.013)	(0.007)	(52.539)
Waves Fixed effect	Yes	Yes	Yes	Yes
Website fixed effect	Yes	Yes	Yes	Yes
Std. err Websites level	cluster	cluster	cluster	cluster
Obs	11238	20556	20556	10967

Notes: Estimates from the DID estimation for all website with advertising before the GDPR. Before GDPR is the omitted group. Column (1) presents estimation for the *Log Number of GDELT URL* dependent variable. Column (2) reports estimation with *Log User Reach per million* as dependent variable. Column (3) presents *Log Page Views per User* as dependent variable. Column (4) reports estimation with *FB average Reaction* as dependent variable. All estimations include waves and website fixed effect.
Standard errors in parentheses and clustered at the website level.
Significance levels: $*p < .10$, $**p < .05$, $***p < .01$.

7 Discussion and Limitations

While our analysis is ongoing, it provides initial insights into the impact of the GDPR on websites' content quantity and users' engagement as a proxy for content quality. Overall, GDPR has reduced the number of third-party cookies and tracking responses, suggesting decreased tracking of users by websites. This decrease is more evident for EU IP addresses visiting US websites, indicating that US websites are taking a conservative approach when dealing with the requirements of GDPR. Furthermore, the enactment of the GDPR may have to some extent negatively affected traffic of EU websites, relative to US sites. However, and importantly, the enactment does not seem to have relatively affected the amount of content that EU websites were able to publish, or the degree of

average social media engagement and interaction with such content.

Before concluding, we feel it is important to highlight some limitations of our analysis. While we are using multiple measures to capture content quantity and quality, they are only proxies that may not fully capture the potential effect of the GDPR. Additionally, while we have classified cookies and http requests to identify tracking and advertising related activity, and devised a way to detect the presence of consent mechanisms, our technical variables are only capturing a part of the technical changes that are possible. For example, there may be different types of consent mechanisms (notices vs granular consent) or websites may choose to move from an ad-supported to a subscription supported model. We are currently working on extensions to measure these possibilities. These additional variables will provide more clarity and precision to our analysis of the effects of the GDPR on online advertising and content publishers.

Finally, despite being over two years into the GDPR, it may still be too early to detect changes in the content produced by publishers. Firms, weighting the cost of compliance against potential fines that may result from enforcement actions, may be inclined to wait until EU authorities provide further clarification on the requirements for compliance. Others still may be justifying data collection and processing under the 'legitimate interest' clause of Article 6. Indeed, a December 2019 report by the Dutch Data Protection Authority found that many popular websites are still placing tracking cookies on the browsers of EU visitors (Authority, 2019). If a significant number of websites are currently not compliant with the GDPR consent requirements, this would make the impact of the regulation on publishers' content weaker and thus more difficult to detect. It's possible that future clarifications or enforcement actions by the EU will trigger smaller scale market shocks as publishers are steered towards compliance in areas such as consent.

8 Conclusion

While previous work has focused on measuring the effects of the GDPR on advertising technologies (such as cookies), the present study attempts to assess the impact of the GDPR on ad-supported content publishers by tracking the potential downstream economic effects of the regulation. We captured a number of metrics related to tracking, traffic, and content variables over several months, both leading up to and immediately following the enforcement of the GDPR.

We examined these variables using descriptive statistics and DID estimations. The DID analysis examined the changes in our variables for US and EU websites viewed from US and EU IP addresses. For websites viewed from the EU relative to websites viewed from the US, our results indicate a reduction in the variables often associated with tracking; we also observed some evidence of a negative impact of the regulation on the traffic of EU websites. However, and importantly, we did not find significant evidence of a negative effect of the regulation on the amount of content that EU websites publish, or the degree of average social media engagement and interaction with such content.

References

Acquisti, A., Taylor, C. and Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*. 54(2), 442–92.

Adjerid, I., Acquisti, A., Telang, R., Padman, R. and Adler-Milstein, J. (2015). The impact of privacy regulation and technology incentives: The case of health information exchanges. *Management Science*. 62(4), 1042–1063.

Anderson, S. and Gabszewicz, J. (2006). The media and advertising: A tale of two-sided markets. In *Handbook of the Economics of Art and Culture*. (p. 568–614.). vol. 1. Elsevier B.V., Amsterdam.

Aridor, G., Che, Y.-K., Nelson, W. and Salz, T. (2020). The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR. *Available at SSRN*.

Article 29 Data Protection Working Party (2017). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*. Retrievable at `https://ec.europa.eu/newsroom/article29/item-detail.cfm?item{_}id=612053`.

Athey, S., Calvano, E. and Gans, J. S. (2018). The impact of consumer multi-homing on advertising markets and media competition. *Management Science*. 64(4), 1574–1590.

Authority, D. D. P. (2019). *AP: veel websites vragen op onjuiste wijze toestemming voor plaatsen tracking cookies*. `https://autoriteitpersoonsgegevens.nl/nl/nieuws/ap-veel-websites-vragen-op-onjuiste-wijze-toestemming-voor-plaatsen-tracking-cookies`.

Cagé, J., Hervé, N. and Viaud, M.-L. (2015). The production of information in an online world. *The Review of Economic Studies*.

Campbell, J., Goldfarb, A. and Tucker, C. (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy*. 24(1), 47–73.

Casadesus-Masanell, R. and Zhu, F. (2013). Business model innovation and competitive imitation: The case of sponsor-based business models. *Strategic Management Journal*. 34(4), 464–482.

Castro, D. (2010). *Stricter privacy regulations for online advertising will harm the free internet*. Technical report. Information Technology and Innovation Foundation.

Choi, J. P., Jeon, D.-S. and Kim, B.-C. (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics*. 173, 113–124.

Dabrowski, A., Merzdovnik, G., Ullrich, J., Sendera, G. and Weippl, E. (2019). Measuring cookies and web privacy in a post-gdpr world. In *International Conference on Passive and Active Network Measurement*. Springer, 258–270.

Davies, J. (2019). *After GDPR, The New York Times cut off ad exchanges in Europe — and kept growing ad revenue*. https://digiday.com/media/gumgumtest-new-york-times-gdpr-cut-off-ad-exchanges-europe-ad-revenue/. Accessed: 2019-05-08.

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F. and Holz, T. (2019). We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Network and Distributed Systems Security (NDSS) Symposium 2019*.

Deloitte (2013). *Economic impact assessment of the proposed General Data Protection Regulation*. Technical report. Deloitte. Retrievable at https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/about-deloitte/deloitte-uk-european-data-protection-tmt.pdf.

Eijk, R. v., Asghari, H., Winter, P. and Narayanan, A. (2019). The impact of user location on cookie notices (inside and outside of the European union). In *Workshop on Technology and Consumer Protection (ConPro'19)*.

Englehardt, S. and Narayanan, A. (2016). Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1388–1401.

Evans, D. S. (2009). The Online Advertising Industry: Economics, Evolution, and Privacy. *Journal of Economic Perspectives*. 23(3), 37–60.

Goldberg, S., Johnson, G. and Shriver, S. (2019). Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic & E-Commerce Outcomes. *Available at SSRN 3421731*.

Goldfarb, A. (2004). Concentration in advertising-supported online markets: An empirical approach. *Econom. Innovation New Tech.* 13(6), 581–594.

Goldfarb, A. (2014). What is different about online advertising? *Review of Industrial Organization*. 44(2), 115–129.

Goldfarb, A. and Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*. 30(3), 389–404.

Goldfarb, A. and Tucker, C. (2012). Privacy and innovation. *Innovation policy and the economy*. 12(1), 65–90.

Goldfarb, A. and Tucker, C. E. (2010). Privacy regulation and online advertising. *Management science*. 57(1), 57–71.

IAB Europe (2017). *Consent*. Retrievable at https://www.iabeurope.eu/wp-content/uploads/2017/11/20171128-Working{_}Paper03{_}Consent.pdf.

IHS Technology (2015). *Paving the way: how online advertising enables the digital economy of the future*. Technical report.

Jia, J., Jin, G. Z. and Wagman, L. (2018). *The Short-Run Effects of GDPR on Technology Venture Investment*. Working paper.

Johnson, G. and Shriver, S. (2019). Privacy & market concentration: Intended & unintended consequences of the GDPR. *Available at SSRN*.

Johnson, G. A., Lewis, R. A. and Nubbemeyer, E. I. (2017). Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research*. 54(6), 867–884.

Johnson, G. A., Shriver, S. K. and Du, S. (2020). Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*.

Lambrecht, A., Goldfarb, A., Bonatti, A., Ghose, A., Goldstein, D. G., Lewis, R., Rao, A., Sahni, N. and Yao, S. (2014). How do firms make money selling digital goods online? *Marketing Letters*. 25(3), 331–341. doi:10.1007/s11002-014-9310-5.

Lefouili, Y. and Toh, Y. L. (2018). Privacy Regulation and Quality Investment. *Working paper*.

Miller, A. R. and Tucker, C. (2009). Privacy Protection and Technology Diffusion: The Case of Electronic Medical Records. *Management Science*. 55(7), 1077–9.

Monic, S. and Feng, Z. (2013). Ad Revenue and Content Commercialization: Evidence from Blogs. *Management Science*. 59(10), 2314–2331.

Ryan, J. (2017). *Can websites use "tracking walls" to force consent under GDPR?* Retrievable at `https://pagefair.com/blog/2017/tracking-walls/`.

Sanchez-Rola, I., Dell'Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vervier, P.-A. and Santos, I. (2019). Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. 340–351.

Shiller, B., Waldfogel, J. and Ryan, J. (2018). The effect of ad blocking on website traffic and quality. *The RAND Journal of Economics*. 49(1), 43–63.

Sørensen, J. and Kosta, S. (2019). Before and after gdpr: The changes in third party presence at public and private european websites. In *The World Wide Web Conference*. 1590–1600.

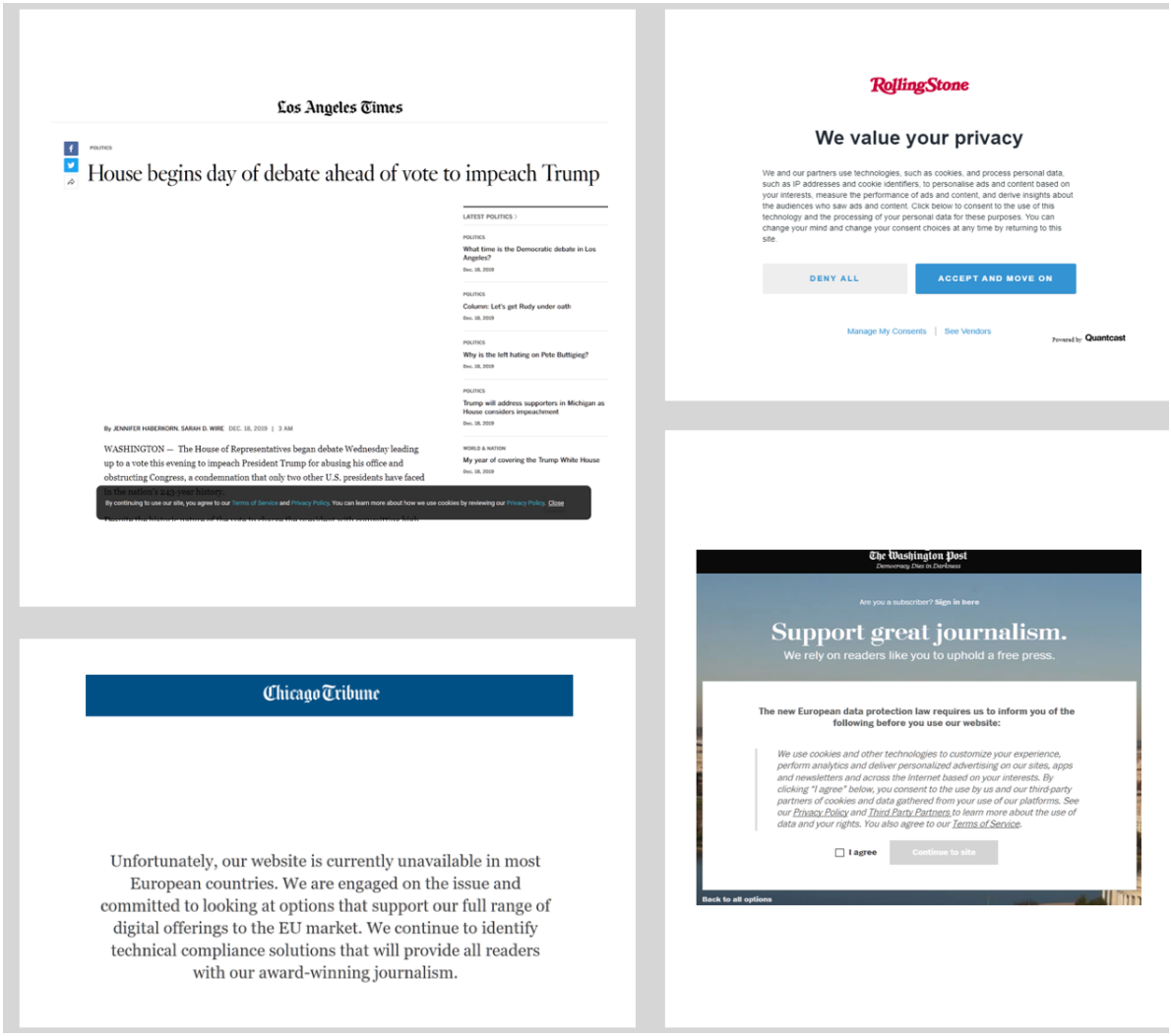
Tucker, C. (2012). The Economics of Advertising and Privacy. *International Journal of Industrial Organization*. 30(7).

UK Information Commissioner's Office (2019). *Update report into adtech and real time bidding*. `https://ico.org.uk/media/about-the-ico/documents/2615156/adtech-real-time-bidding-report-201906.pdf`.

Urban, T., Tatang, D., Degeling, M., Holz, T. and Pohlmann, N. (2020). Measuring the Impact of the GDPR on Data Sharing in Ad Networks.

9 Annex

Fig. 8 *Types of Consent Mechanisms*



Note: This figure presents different kinds of consent mechanisms

Table 10: Description of the variables and identification methodologies

Variable	Description	Information
Content Variables		
GDELT URLs	Number of URL collected from GDELT	"GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world" GDELT
Reach per Million	Average number of users visiting a domain per million random users on the internet	From alexa website
Rank	Rank computed by Alexa based on three months aggregated historical traffic from Alexa toolbar users and page views	
Page views per million	The fraction of all the page views by toolbar users that go to a particular websites	
Page views per user	Page views measure the number of pages viewed by Alexa Toolbar users	
Technical Variables		
Persistent Cookies	Cookies staying more than 30 days on browser	Identify by the author
Session Cookies	Session cookie contains information that is stored in a temporary memory location and then subsequently deleted after the session is completed or the web browser is closed	From Open WPM
3rd party	From an external domain	Identify by the author
1st party	From the domain itself	Identify by the author
Consent mechanism	Indicate the presence of a Consent mechanism on the website	Using Blocklist I don't care about cookies
Ads - Disconnect	Technical variable lag as Advertising by Disconnect	"A tracker which also displays ads or marketing offers. These types of ads can track your personal information and expose you to malware, even if you don't interact with them." from Disconnect link
Ads - Easylist	Technical variable flag as Advertising by Adlock Easylist	easylist link
Social - Disconnect	Technical variable flag as Social networking by Disconnect	"A tracker may be classified as social if it uses tracking techniques that allow a social networking service to track your web browsing activities even when you are not on the social network's website or app". From Disconnect link
Analytics	Technical variable flag as Analytics by Disconnect	"A tracker which collects your information and may build a profile based on your online activity that can be connected with your real name or other unique identifier". From Disconnect link
Google/Facebook/Twitter - Disconnect	Identification of exchange between the browser and the 3rd parties Google, Facebook or Twitter by Disconnect	
Content - Disconnect	Technical variable flag as Social networking by Disconnect	"Third parties that are critical to the delivery of content to a web page as blocking them could cause the webpage to "break"". From Disconnect link
Adblockfilter	Technical variable flag as blocker of Adblocker from Easylist	easylist link
Tracking variable	Technical variable flag as Privacy by Easylist	
Variable from Country specifics list	Cookies flag as Country specific variable from Easylist, meanng a specific language (French, German, Farsi ...)	