Privacy Regulations and Online Search Friction: Evidence from GDPR

Yu Zhao   Pinar Yildirim   Pradeep Chintagunta*

This version: August 2021

Abstract

How do privacy regulations in the market impact online search for products and information? This paper investigates the impact of the General Data Protection Regulation (GDPR for short) on consumers' online browsing and search behavior using consumer panels from four countries, United Kingdom, Spain, United States, and Brazil. We find that after GDPR, a panelist exposed to GDPR submits 21.6% more search terms to access information and browses 16.3% more pages to access consumer goods and services compared to a non-exposed panelist, indicating higher friction in online search. The implications of increased friction are heterogeneous across firms: Bigger e-commerce firms see an increase in consumer traffic and more online transactions. The increase in the number of transactions at large websites is about 6 times the increase experienced by smaller firms. Overall, the post-GDPR online environment may be less competitive for online retailers and may be more difficult for EU consumers to navigate through.

Keywords: General Data Protection Regulation, online privacy, consumer search, e-commerce

*Zhao is a doctoral candidate at Wharton School of the University of Pennsylvania, e-mail: yzhao25@wharton.upenn.edu. Yildirim is Assistant Professor of Marketing, Marketing Department, The Wharton School, University of Pennsylvania, Philadelphia PA 19104, (215) 746 2369, pyild@wharton.upenn.edu. Chintagunta is Joseph T. and Bernice S. Lewis Distinguished Service Professor of Marketing, Booth School of Business, University of Chicago. 5807 South Woodlawn Avenue, Chicago, Illinois 60637, (773) 702 8015, pradeep.chintagunta@chicagobooth.edu. Authors thank Netquest for providing the data that made this paper possible, and Tesary Lin, Christophe Van den Bulte, Walter Zhang, and the audiences at the 2020 Marketing Science Conference and Theory and Practice Conference for their valuable feedback. The authors thank the Marketing Science Institute for their generous funding for this project. Yildirim thanks Wharton School's Mack Institute and the Dean's Research Grant for their financial support. Chintagunta thanks the Kilts Center at Chicago Booth, University of Chicago for financial support. All errors are our own. All correspondence regarding the manuscript can be sent to Yildirim.

1 Introduction

On May 25th, 2018, the European Union (EU) implemented a series of laws which regulate the practice of collecting, storing, and using consumer data for companies that serve consumers located in the EU region. Referred to as the General Data Protection Regulation, or GDPR for short, these regulations extend the scope of previously existing consumer privacy protections and introduce new mandates by firms utilizing consumer data (Council of European Union, 2014). GDPR requires informed, opt-in consent from customers prior to data collection and gives consumers the right to access, correct, and erase their personal data. Simultaneously, GDPR requires firms to take proactive steps to anonymize and secure personal data by developing protocols to respond to individual data requests in a timely fashion and appoint a data protection officer to oversee compliance activities. Failure to comply with GDPR can lead to fines up to 4% of the overall firm revenues. GDPR thus creates a barrier for firms wishing to take advantage of consumer data in their marketing activities, e.g., to send firm communications, to acquire new consumers, or to target consumers with advertising. These implied frictions can impact consumers' efforts to access information and products in the digital environment. In particular, they could make online search costlier and potentially alter search outcomes.

In this study, we estimate the impact of online privacy regulations on search for content and products by studying the implications of GDPR for consumers. These implications are hard to predict, ex-ante. On the one hand, GDPR offers privacy benefits, i.e., protections to consumers. Existing studies show that consumers respond positively to privacy policies set by firms (Tsai et al., 2011) and dislike sellers that use their personal information to target them in their ads (Goldfarb and Tucker, 2011). Reduced violations of privacy may help consumers feel safe when engaging in online activities, enabling them to browse and transact with more confidence. On the other hand, GDPR introduces costs for firms to collect and utilize consumer data, adding to informational friction in online environments. As a firm not only faces higher costs but also has a lower ability to use consumer data in its marketing communications, it may fail to deliver content and product information to consumers efficiently. This inefficiency may, in turn, hurt consumers as they may increase their search effort, and presumably face worse search outcomes. It is therefore important to ask if the benefits of GDPR due to enhanced privacy make up for the losses from increased informational friction.

We examine the net effect of these two contrasting consumer outcomes using extensive online browsing and search data, with panelists from four different countries in and outside the EU region—UK, Spain, US, and Brazil. We identify the causal impact of GDPR on consumer online browsing and search using a difference-in-differences approach, exploiting the geographical reach of GDPR: GDPR protects consumers located in the EU region (Spain and UK panels) but not those beyond (Brazil and US panels). We estimate changes in consumer browsing and search behaviors after GDPR by comparing EU and non-EU consumers. On the firms' side, we also estimate the changes in website traffic by comparing websites with higher and lower pre-GDPR EU user penetration. Despite GDPR specifying protection only for consumers located within the EU region, there could be spillover effect to countries outside of EU as websites serving EU customers are subject to GDPR requirements thus may update their privacy policies and upgrade their technologies regarding data collection and anonymization. As a result, our results reveal a relative effect of GDPR on protected consumers relative to consumers beyond GDPR's scope, or panelists in our control groups: the US and Brazil panels.

After GDPR, EU panelists in our data increased their online activities in total and per domain. An EU consumer on average visits 14.9% additional domains, browses 0.39% more pages on a domain, and spends 44.7% more time on the internet after GDPR goes into effect compared to the non-EU consumers. These increased engagement outcomes from consumers are consistent with both the enhanced privacy benefits of GDPR and the inefficiency firms face to reach out to customers.

To further investigate whether there is a change in the level of frictions, we then estimate the impact of GDPR on consumer search. We focus on two types of search: (1) search for general information by submitting search terms to a search engine or browser, and (2) search for product information by browsing products on e-commerce sites. For the first type, we utilize a novel dataset of consumer keywords along with natural language processing methods to identify general information search episodes. For the second type, we parse the URLs consumers visit to identify the products they look for. The comparison of EU and non-EU panelists before and after GDPR demonstrates that, keeping the topic fixed, the search effort, measured by number of search term submitted for the same topic, increased by 4.8% for EU panelists relative to their non-EU peers after GDPR, consistent with the idea of higher information friction. When searching for products, EU-panelists spent 11.2% more time browsing products, considered 10.6% additional products in 6.2% more unique e-commerce sites relative to non-EU panelists. These findings are consistent with higher friction to find products. In contrast to these findings, the search effort was shorter for EU panelists than their non-EU peers when a search resulted in a transaction. Among the possible mechanisms at play here is consumers' selection into buying from known alternatives, which is further supported when we test for the heterogeneity of the GDPR effect across domains. We find that smaller websites experience significant declines in their traffic, but not large firms. This implies that the negative effects of GDPR are felt disproportionately more by the smaller firms. The number of checkouts increases for larger firms, and their increases are about 6 times of the increase for smaller e-commerce sites.

For policymakers, our findings imply that privacy policies may be associated with an increase in consumer search effort online, for both general information and product-related searches. The implementation of the policy also coincides with an exacerbated inequality between larger and smaller businesses. We find a decrease in the consideration set size - the number of e-commerce sites examined - when consumers search for products before buying, and the total number of checkouts increased for larger e-commerce sites after GDPR. These findings suggest there could be more concentration in the online commerce environment, which is contrary to the policy's original intention.

Our study contributes to the growing literature on consumer privacy and the impact of privacy regulations(e.g., Lin, 2020; Acquisti et al., 2015; Johnson et al., 2020; Ke and Sudhir, 2020). Goldfarb and Tucker (2011) document that privacy regulations in the EU resulted in reduced ad effectiveness, as we also argue in our paper. More recently, a number of studies focused on the implications of GDPR, in particular, GDPR's impact on the entry and exit of new EU-based ventures (Jia et al., 2021) and entry of new mobile apps (Janssen et al., 2021), on the interconnections between technology providers (Peukert et al., 2020) and concentration of third party technology vendors (Johnson and Shriver, 2021; Batikas et al., 2020), and on content production (Lefrere et al., 2020).

Two studies focusing on consumer response to GDPR are particularly relevant to ours. Aridor et al. (2020), using data from an intermediary in the travel sector, find that after GDPR, fewer consumers opt in to share their data, but for consumers that still share data, their behavior becomes more predictable. Goldberg et al. (2021), using data from online firms which utilize Adobe's website analytics tools, document a decline in users' pageviews, opposite to what we find. Since they highlight the challenges firms face to collect data from consumers after GDPR, Aridor et al. (2020) and Goldberg et al. (2021) are complementary to ours. At the same time, our study has a number of advantages and differentiating points. First, they face a selection problem due to not observing consumers who opt out from data collection after GDPR goes into effect. Our study does not suffer from this issue, as we work with a consumer panel with little attrition. Second, differently from Aridor et al. (2020) and Goldberg et al. (2021), our data do not come from a single industry or a single intermediary, which may introduce a selection issue. We work with a panel that is chosen to represent the broad characteristics of the national population and records all online activities of users at the URL level. Our analysis takes advantage of the panel nature of our data to strengthen the causal identification.

This study also contributes to the literature on search (e.g., De los Santos et al., 2012; Bronnenberg et al., 2016; Seiler and Pinna, 2017; Yavorsky et al., 2021), where frictions resulting from increased costs of search are well documented in theoretical (e.g., Stigler, 1961; Diamond, 1971) and empirical consumer search literature (e.g., Sorensen, 2000; Kim et al., 2010). Our paper contributes to this field by documenting the search implications of privacy policies and jointly identifying consumer search effort and scope of search, using URLs and text analysis.

Finally, our study may also be of relevance to the application of natural language processing methods on processing consumer data. Given the growing interest among marketers on using machine learning to process consumer data (Archak et al., 2011; Liu and Toubia, 2018; Timoshenko and Hauser, 2019), our study may be of relevance to text processing literature as well.

In the rest of the paper, we proceed in the following way. We introduce the data sets we use in section 2. In section 3, we discuss our empirical specifications and results of GDPR's effect on consumer browsing and search activity. Section 4 focuses on the heterogeneous effects of GDPR and section 5 carries out a set of robustness checks. In section 6, we conclude.

2 Data

Consumer Browsing We use data from Netquest, a consumer insights company that tracks individuals' online browsing activities in a number of countries around the world. Our clickstream data includes browsing panels from four different countries in and outside of EU: UK, Spain, US, and Brazil. The data set covers the period starting from January 31st to September 1st, 2018. This period coincides with the implementation of GDPR on May 25th, 2018. The clickstream data set includes the panelist identifier, date and time of visiting a website, full URL of the visited site thus the domain name, and the time spent at the visited URL.[1] To eliminate cases where a visitor may accidentally open a window or may leave a screen open without actively browsing, we drop clicks where the page view is shorter than 2 seconds or longer than 12 hours.[2] In total, there are over 8 million observations of visits made by 6,000 panelists (1,500 in each of the four countries) to 887,525 domains throughout the 34 weeks. We convert the clickstream data into three balanced panels at user-week level, website-week level, and user-website-week level. The ability to track individuals over time, i.e., the panel structure of our data, is crucial to our analysis. By controlling for individual-level activities or including individual fixed effects in our analysis, we are able to rule out alternate explanations such as panelists from the EU region having a different level of online activities, confounding with GDPR's effect.

Table 1 presents summary statistics of the user-week panel, breaking the data by user-region (EU vs. non-EU) and by visits before and after GDPR. We note two observations from the summary. First, EU panelists on average are less active online than their non-EU peers. Second, for both EU and non-EU users, there are declines in average number of domains visited, average total time spent online, and per-page view time after GDPR, but the decline for an EU user is smaller than that of a non-EU user: the average decline in non-EU users' number of unique domains is 4.51, and the decline for EU users is 2.99 less. Similar patterns hold for averages of total time spent online as well as per-page view time.

In Table 2, we present summaries from the user-website-week panel: total time a panelist spends on a domain in a week, number of pages viewed at this domain, and per-page

[1] Compared to other well-known consumer browsing panel data sets (e.g., ComScore), a key advantage of the Netquest panel is that it includes the full URL, excluding identity-revealing information, which allows us to extract information about the activity of users such as their online e-commerce browsing and transaction sequence.

[2] Corresponding to 5th and 99th percentiles, respectively.

Table 1: Summary statistics: Consumer weekly browsing behavior, by EU and GDPR

	Before GDPR		After GDPR	
	non-EU users	EU users	non-EU users	EU users
No. unique domains	63.09 (60.37)	53.20 (55.11)	58.58 (58.47)	50.99 (53.20)
Total time online (sec)	58,690.45 (58,402.84)	37,998.83 (44,089.09)	56,297.57 (59,127.00)	37,439.34 (44750.60)
Per-page view time (sec)	46.56 (43.67)	38.80 (31.06)	47.52 (40.28)	39.71 (34.56)
No. obs	44,421	43,928	49,647	49,096

Notes: The summary statistics are computed across users and week. An observation is a user-week. For number of unique domains and total time spent online, the average and the standard deviations are calculated by assigning the value zero when a panelist is absent in a week. For per-page view time, the average and the standard deviations are computed by excluding the panelists who were not online.

average view time on this domain. These variables provide information on the intensive margin of consumption within a domain, and we use these variables to study how consumer behaviors change for a given domain after GDPR. We also note that there is a high degree of heterogeneity across users, domains, and time, as we find large standard deviations across observations in Table 2.

In Appendix B.6, we present the summary of panelists' demographics (age, gender, education, monthly income, and family size) for each of the four countries. While the numbers look comparable, we expect some differences to exist because country-level population statistics are likely to be different from the statistics of the internet-using population. Moreover, our panel consists of an adult population giving consent to share their data.

App Usage on Mobile Devices A second data set from Netquest contains consumers' usage of mobile apps on devices such as smartphones and tablets. An observation in this data set is an app session. For each app session, we observe the panelist ID, the name of the app being accessed, the type of device—a smartphone or a tablet— that was used to access the app, the operating system of the device, connection type, and finally the duration of app usage, measured in minutes. In total, there are 51,864 unique apps (names) accessed. On average, an app is accessed by 4.16 unique users. We dropped the sessions that lasted less than 1 minute or longer than 720 minutes. We present summary statistics of mobile app access in Table 3, broken down by panelist region and before/after GDPR. As the mobile

Table 2: Summary statistics: Consumer browsing activities within a domain, by EU and GDPR

	Before GDPR		After GDPR	
	non-EU users	EU users	non-EU users	EU users
Total time (sec)	69.77 (1,412.44)	49.08 (1,050.73)	66.95 (1,437.45)	48.36 (1,066.70)
No. pages viewed	1.90 (31.99)	1.52 (27.40)	1.81 (31.71)	1.46 (26.18)
Per-page time (sec)	2.82 (23.43)	2.34 (20.25)	2.63 (22.63)	2.30 (21.19)
No. obs	37,249,448	33,956,633	41,631,736	37,951,531

Notes: The summary statistics are computed from a balanced panel such that total time spent on a domain, pages viewed on a domain are filled with zeroes when a panelist does not visit a domain in a week. Average per-page time is calculated from dividing total time spent by total number of pages viewed on a domain in a week, and it is zero if a panelist does not visit a domain in a week.

app industry heavily uses consumer personal information for ad delivering and tracking, we expect GDPR to impact consumer app usage as well. To this end, we use this data set to examine whether GDPR is associated with changes in breadth and intensity of consumer app access and usage.

Search Keywords A third data set from Netquest contains search keywords entered into a search engine or a browser. Each observation in this data set is a search term, which is usually a short sentence or a multi-word phrase. For each search term, we observe the ID of the panelist who submitted the search term, timestamp of submission, and the search engine used. After dropping search terms with less than the first and higher than the 99th percentile of the word lengths for each country, an average search term contains 5.07 words in the US panel, and 4.64 words from the UK panel. We also observe consumers' search terms related to GDPR. With these keywords, we can also look at consumers' search of the keyword "GDPR", which is evident in Figure B.2 in Appendix B.3.

Table 3: Summary statistics: Consumer app usage, by EU and GDPR

	Before GDPR		After GDPR	
	non-EU users	EU users	non-EU users	EU users
Total session time (minutes)	815.77	575.08	842.46	615.38
	(1080.28)	(837.64)	(1082.08)	(851.22)
Number of apps	15.08	10.04	15.07	10.31
	(17.16)	(12.67)	(16.65)	(12.48)
Avg. per-app time (minutes)	38.35	38.30	40.48	42.04
	(45.26)	(52.55)	(46.24)	(54.24)
No. obs	129,608	125,205	144,856	139,935

Notes: The summary statistics are computed from a balanced panel such that total time spent and number of apps accessed each week are zeroes when a user-device-connection-operating system combination is not observed in a week. An observation is a user-device-connection-operating system-week. Total session time is the sum of all sessions time within the same week, for the same panelist, on the same device (smartphone or tablet), under the same connection (3G, wifi, or unknown), using the same operating system (Android or iOS). Average per-app time is calculated by dividing the total time of the sessions in a given week by the total number of apps used in the same week, and it is zero if number of apps accessed is zero.

3 Empirical Analysis

3.1 Change in Consumer Browsing

We analyze the changes in consumers' use of the internet by focusing on the breadth and intensity of browsing. We employ a difference-in-differences (DID from here on) approach with controls to compare pre-GDPR and post-GDPR outcomes for EU and non-EU panelists. In general, we run a DID specification as given in Equation 1:

$$log(Y_{it}) = \alpha_1 + \alpha_2 GDPR_t \times \text{EU}_i + \alpha_3 \text{EU}_i + \tau_t + \epsilon_{it} \quad (1)$$

where Y_{it} stands for a set of outcomes of consumer browsing activity, including (1) the number of unique domains visited, (2) the total time spent online (in seconds), and (3) per-page view time (in seconds) by panelist i in week t. We log-transform the outcome variable to account for any skewness in the distributions. EU_i is a dummy indicating that panelist i is from EU region (i.e., UK or Spain) or not (i.e., US or Brazil). GDPR_t indicates if the week is on or after the week of May 25th, 2018, and takes the value zero otherwise. Here, we are interested in the sign of α_2 which is the change in the outcome variable for the EU users after GDPR relative to non-EU users. We include week fixed effects, τ_t, to account for events that create temporary fluctuations in browsing. We carry out the analysis on a balanced panel. In some of our analysis presented later, we also control for panelist fixed effects to account for the time-invariant characteristics of users, which may make them more or less prone to being online and browse more or fewer webpages.

We illustrate the ambiguity in GDPR effects with a simple theoretical model in Appendix A. The identification of the GDPR's impact on Y_{it}, α_2, is based on the observation that GDPR provides protection for users inside the EU region but not for those outside. The non-EU panelists (namely, our panelists from Brazil and the US) serve as the control group in this specification. This strategy assumes that in the absence of GDPR, EU panelists and non-EU panelists would have similar browsing patterns, or that the parallel trends assumption would hold. We verify this assumption in Appendix C: starting the week prior to the official GDPR date, browsing behavior starts to significantly differ between the EU and non-EU panels.

This identification strategy ultimately estimates the difference in GDPR's effect on EU and non-EU panelists. It is possible that non-EU panelists are also impacted by GDPR, for instance, if non-EU panelists are visiting domains which serve predominantly EU users, or if sites adopt blanket privacy policies regardless of the IP addresses of the users. We anticipate that the exposure of non-EU panelists to GDPR and firms that are subject to them will be lesser, since these domains are not required by law to offer protections for non-EU residents. However, given the fact that some non-EU panelists will be 'treated,' the impact of GDPR we estimate should be read as the differential impact of GDPR on the EU-panelists relative to the non-EU panelists, and all coefficients should be interpreted accordingly.

Table 4 summarizes the results from specification (1). Columns (1) and (2) show the results for the number of unique domains visited in a week, columns (3) and (4) summarize the results for total time (in seconds) panelist spends online, and columns (5) and (6) are the results for per-page view time, measured in seconds, with week and/or user fixed effects. In all columns, the differential impact of GDPR on EU users relative to non-EU panelists is positive and significant at 1% level. The coefficient for the number of unique domains corresponds to an increase of approximately 14.9%, the coefficient for total time spent corresponds to a 44.8% increase, and the coefficient for per-page view time corresponds to a 15.6% increase

Table 4: Impact of GDPR on consumer browsing

	log(no. unique domains)		log(total time (seconds))		log(avg. per-page time (seconds))	
	(1)	(2)	(3)	(4)	(5)	(6)
GDPR × EU	0.139***	0.139***	0.370***	0.370***	0.145***	0.145***
	(0.013)	(0.010)	(0.031)	(0.025)	(0.011)	(0.009)
EU	-0.300***		-0.926***		-0.327***	
	(0.010)		(0.022)		(0.008)	
Constant	3.484***	3.335***	9.549***	9.089***	3.361***	3.199***
	(0.005)	(0.004)	(0.011)	(0.009)	(0.004)	(0.003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE		Yes		Yes		Yes
No. obs	187,092	187,092	187,092	187,092	187,092	187,092
Adjusted R^2	0.039	0.46	0.045	0.36	0.033	0.32
Mean of DV	56.396	56.396	47,620.58	47,620.58	38.721	38.721

Notes: The specification is $log(Y_{it}) = \alpha_1 + \alpha_2 GDPR_t \times \text{EU}_i + \alpha_3\text{EU}_i + \tau_t + \epsilon_{it}$. Each observation is a user-week record, analysis done at user-week level. Y_{it} is the number of unique domains visited by panelist i in week t, and total time panelist spends online in a week. EU_i is an indicator of whether a user is from the EU region - namely, from UK or Spain panel. $GDPR_t$ equals 0 if week t is before the week of May 25th, 2018; it equals 1 if week t is the week of May 25th, 2018 or beyond. Heteroscedasticity-robust standard errors clustered at the panelist level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

in per-page view time.[3] [4] [5]

3.1.1 Change in Consumer App Usage

Complementing the previous results, we also investigate how consumers' mobile app usage has changed. We employ a DID approach and compare pre-GDPR and post-GDPR outcomes for EU and non-EU panelists. We run the following specification:

$$log(Y_{ikgpt}) = \alpha_1 + \alpha_2\text{GDPR}_t \times \text{EU}_i + \alpha_3\text{EU}_i + \tau_t + \xi_k + \zeta_g + \nu_p + \epsilon_{ikgpt} \tag{2}$$

[3] In Figure 3, Section 5, we show that these results are robust to small modifications to window sizes.

[4] There may be a concern about whether it is the fact that GDPR goes into effect that results in an increase in the time consumers spend online. For instance, a pop-up screen asking for consent can divert consumers' attention. Table C.10. in the appendix runs an alternative version of Table 4, removing the first click to any page, assuming that the first clicks may be asking for consent or posting GDPR notices. Our results remain qualitatively unchanged and quantitatively similar with this robustness exercise. For more details, please see Appendix C.7.

[5] We also assess whether the increase in time spent online is due to an increase in time spent in a single browsing session or the number of browsing sessions, or both. We do not find a significant increase in within-session browsing time, but there is a significant increase in the number of sessions browsed per week has increased. For more details, please see Appendix C.7, Table C.9.

Table 5: Effect of GDPR on consumer mobile app usage

	log(no. apps)		log(total time (minutes))		log(per-app time (minutes))	
	(1)	(2)	(3)	(4)	(5)	(6)
GDPR × EU	0.033***	0.033***	0.104***	0.104***	0.076***	0.076***
	(0.007)	(0.006)	(0.015)	(0.014)	(0.010)	(0.009)
EU	-0.265***		-0.411***		-0.124***	
	(0.005)		(0.011)		(0.007)	
Constant	1.920***	1.790***	4.478***	4.276***	2.657***	2.596***
	(0.002)	(0.002)	(0.005)	(0.005)	(0.003)	(0.003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Connection FE	Yes	Yes	Yes	Yes	Yes	Yes
Device FE	Yes	Yes	Yes	Yes	Yes	Yes
Operating System FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE		Yes		Yes		Yes
No. obs	539,604	539,604	539,604	539,604	539,604	539,604
Adjusted R^2	0.26	0.45	0.19	0.37	0.11	0.28
Mean of DV	12.673	12.673	715.122	715.122	39.865	39.865

Notes: The specification is $log(Y_{ikgpt}) = \alpha_1 + \alpha_2 \text{GDPR}_t \times \text{EU}_i + \alpha_3 \text{EU}_i + \tau_t + \xi_k + \zeta_g + \nu_p + \epsilon_{ikgpt}$. Each observation is a record at user-device-connection-operating system-week level. The outcomes are total session time (in minutes) (columns (1) and (2)), total number of apps accessed (columns (3) and (4)), and finally the average time spent on an app (in minutes) (columns (5) and (6)). EU_i is an indicator of whether a user is from the EU region - namely, from UK or Spain panel. $GDPR_t$ equals 0 if week t is before the week of May 25th, 2018; and 1 otherwise. Heteroscedasticity-robust standard errors clustered at the user-device-connection-operating system level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

where Y_{ikgpt} is the outcome of panelist i, using device k with operating system g with connection p in week t. We focus on three outcomes: the total session time (in minutes) (columns (1) and (2)), the total number of apps accessed (columns (3) and (4)), and finally the average time spent on an app (in minutes) (columns (5) and (6)). We control for the device (tablets or smartphones), operating system (Android or iOS), connection (3G or Wi-Fi), panelist, and week fixed effects.

Table 5 demonstrates significant estimates of $\text{GDPR}_t \times \text{EU}_i$ in all columns under different specifications for all three outcomes. In column (2) the estimate of the interaction term is 0.033, which implies a 3.4% increase in the number of apps accessed per week. In column (4), the estimate is 0.104, which implies GDPR is associated with a 10.9% increase in total time spent on the mobile apps compared to the pre-GDPR mean; for average per-app use time, the estimate in column (6) is 0.076, i.e., GDPR is associated with a 7.9% increase. These results show a pattern consistent with our previous findings: consumers have increased activity in both desktop browsing and mobile app usage. This finding is somewhat different from

Janssen et al. (2021), where the total number of new app entries declined. One explanation is that our analysis is based on tracking the consumer's activities over time, and consumers can still use a larger number of apps after GDPR, even though the total number of new apps declines. Further, it is possible that there is a distributional shift of app usage, i.e., the larger, more frequently downloaded apps attract more new visitors after GDPR, while smaller ones have fewer new visitors after GDPR. This is indeed the case for websites. We provide details on the heterogeneity effect of GDPR across websites in section 4.

3.1.2 User Behavior at the Domain Level

The previous analysis provides a summary of how consumers' *overall* browsing activity has changed after GDPR. In this section, we take a closer look at how consumer browsing activity has changed for a *given* domain after a domain's compliance. For this purpose, we collect domain-level GDPR policy update times.

For each domain in the Netquest data, we scraped its posted privacy policy from its website in 2019 by searching for the links containing terms "privacy policy," "user terms," "terms and policy," "cookie policy," and "legal terms" on that domain's landing page, we then scraped the text on these pages.[6] We obtained update times indicating GDPR compliance for 14,551 websites. If a site's privacy policy did not mention GDPR explicitly or did not indicate a date for updating policies, we assigned the official GDPR date (May 25th, 2018) as the policy change date, since after this date websites are subject to penalties if they fail to comply. We check robustness of our results, assuming a uniform policy enforcement time in Appendix C.7, and find consistent results. Figure B.1 shows the distribution of GDPR policy update dates in this data set. Majority of the update dates are around the official GDPR date of May 25, 2018.

We investigate if a panelist's browsing behavior changed for a *given* site, after the website's implementation of GDPR in specification (3):

$$log(Y_{ijt}) = \gamma_0 + \gamma_1 \text{GDPR}_{jt} + \gamma_2 \text{GDPR}_{jt} \times \text{EU}_i + \theta_i + \xi_j + \tau_t + \epsilon_{ijt} \tag{3}$$

[6] In the scraped text, we searched for a mention of a GDPR-related policy update term and update date. GDPR-compliance is indicated by the mention of keywords "GDPR," "general data protection regulation," "data controller," "data protection officer," and "regulation 2016/679," We obtain each site's policy change date by locating phrases in its policy page such as "updated at/on," "last modified at/on" or "last updated at/on," and extracted dates of the time the policy was last modified.

where Y_{ijt} corresponds to one of three outcome variables of interest: (1) average per-page view time (in seconds), (2) total time spent on the site (in seconds), and (3) the number of pages clicked on by panelist i on site j in week t. GDPR_{jt} here is a website-week level indicator that takes the value 1 if the site has already adopted GDPR-relevant policies by week t, and zero otherwise. In the analysis we control for week (τ_t), panelist (θ_i) and domain (ξ_j) fixed effects. We are interested in the coefficient γ_2, which estimates the differential activity of the EU users relative to non-EU users on website j, after site j's adoption of GDPR policies.

In Table 6, columns (1)–(3) report the results for the number of pages visited for a given domain. The coefficient for $\text{GDPR}_{jt} \times \text{EU}_i$ is 0.0039 (column (3)), and is significant at 1% level. The columns (4) – (6) report the results of total time spent on a domain where $\text{GDPR}_{jt} \times \text{EU}_i$ has a coefficient of 0.011 (column (6)). Finally, columns (7)-(9) report the results related to duration spent per page of a domain (in seconds) where the coefficient of the interaction term is 0.0078 (column (9)). All estimates, with various fixed effect specifications, are significant at the 1% confidence level and point to an increase in the activity of EU users relative to the non-EU users. The magnitudes are robust to the inclusion of week, panelist, and domain fixed effects. The results altogether show that, on average, the EU panelists browsed more pages, spent more time on a domain and each page of a domain after a domain adopted GDPR-related policies. The number of pages viewed per domain increased by 0.39% (columns (3)), total time spent on a domain increased by 1.11% (columns(6)), and the time spent on a page increased by 0.75% (columns (9)).

The increase in the overall browsing behavior is as expected, but the drivers are not clear. It is consistent with "enhanced privacy protection" increasing consumers' desire to visit new domains and spend more time online due to increased data privacy. It is also consistent with "higher information friction." The post-GDPR environment is costlier for firms in terms of collecting and utilizing consumer data for marketing – including targeted ads and promotional emails, and firms' ability to track consumers and deliver tailored advertisements deteriorates. Therefore, it is also more costly for consumers to search and find information or products and services they need. The latter mechanism is consistent with the expectation that consumers exert more search effort to obtain the same type of information or on the same types of products when compared to pre-GDPR regime. We will assess these ideas in Section 3.2.

Table 6: Impact of GDPR on consumer browsing activities on a domain

	log(no. pages viewed)			log(total time (seconds))			log(avg per-page time (seconds))		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GDPR (domain) × EU (user)	0.0038***	0.0038***	0.0039***	0.011***	0.011***	0.011***	0.0078***	0.0076***	0.0075***
	(0.00019)	(0.00019)	(0.00017)	(0.00041)	(0.00041)	(0.00038)	(0.00026)	(0.00026)	(0.00025)
GDPR (domain)	0.024***	0.026***	-0.0012**	0.046***	0.052***	-0.00089	0.024***	0.028***	-0.00022
	(0.00049)	(0.00049)	(0.00061)	(0.0011)	(0.0011)	(0.0014)	(0.00071)	(0.00071)	(0.00091)
EU (user)	-0.016***			-0.038***			-0.023***		
	(0.00014)			(0.0003)			(0.00019)		
Constant	0.13***	0.12***	0.13***	0.31***	0.29***	0.31***	0.21***	0.19***	0.21***
	(0.00026)	(0.00025)	(0.00032)	(0.00058)	(0.00056)	(0.00071)	(0.00038)	(0.00037)	(0.00048)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panelist FE		Yes	Yes		Yes	Yes		Yes	Yes
Domain FE			Yes			Yes			Yes
No. obs	150,789,348	150,789,348	150,789,348	150,789,348	150,789,348	150,789,348	150,789,348	150,789,348	150,789,348
Adjusted R^2	0.0011	0.0057	0.18	0.0013	0.0062	0.14	0.0014	0.0061	0.091
Mean of DV	1.68	1.68	1.68	58.95	58.95	58.95	2.53	2.53	2.53

Notes: The table presents results to the specification $log(Y_{ijt}) = \gamma_0 + \gamma_1\text{EU}_i + \gamma_2\text{GDPR}_{jt} + \gamma_3\text{GDPR}_{j,t} \times \text{EU}_i + \tau_t + \theta_{ij} + \epsilon_{ijt}$ where the dependent variables are (logged) number of pages viewed (columns (1)-(3)), total time (seconds) spent (columns (4)-(6)), and total number of pages clicked on a domain (columns (7)-(9)). EU_i is a dummy indicating whether panelist i is from the EU region, i.e., from the UK or Spain panels. $GDPR_{jt}$ is a dummy which equals to 1 if week t is after the week when website j updated its privacy policy in compliance with GDPR; we control for week, panelist, and domain fixed effects. We fill the panel so that for a domain-user combination, if it is missing in a week, we assign 0 to that week's outcomes. Analysis is done at domain-user-week level. In total, we have 150,789,348 observations where observations are at user-domain-week level. Heteroscedasticity-robust standard errors clustered at domain-user level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

3.2 Change in Consumer Search

In the previous section, we documented that EU consumers' online activity increased after GDPR compared to their non-EU peers and laid out the two possible mechanisms consistent with these results. In this section, we further test these mechanisms by examining if the effort consumers need to exert to find information has changed after GDPR. Informational frictions can alter the effort that consumers put into search for various topics such as news, educational materials, entertainment, healthcare, art, and other content, as well as for goods and services. To test the change in effort, we will run variations of the following specification:

$$\text{Search\_effort}_{ikt} = \gamma_0 + \gamma_1 \text{UK}_i + \gamma_2 \text{GDPR}_t \times \text{UK}_i + \theta_i + \eta_t + \nu_k + \varepsilon_{ikt} \tag{4}$$

where $\text{Search\_effort}_{ikt}$ indicates various measures of search effort of panelist i related to topic or product category k in week t. UK_i is a dummy which equals 1 if panelist i is from the UK panel.[7] We also control for category/topic, individual fixed effects and week fixed effects as before.

Ideally, to measure a consumer's search effort, one would need to track consumers' activities like search terms submitted to a search engine, and product pages browsed before a terminal action, such as a visit to a domain or a transaction. As our data includes the search terms panelists submit and URLs of the clicked pages, we measure search effort exerted on a topic by the number of search terms submitted relating to that topic, and we measure effort on product search using page views in the same product category. In what comes next, we will first focus on the search for topics and then the search for products and services, then test if consumer search is altered by GDPR.

3.2.1 Search for Information

To test if consumers' effort to access topics changed after GDPR, we identify a set of latent search topics. Consumers look for information on these topics by typing search keywords in their browser or search engine. We take advantage of the search terms data to identify latent topics, relying on the fact that panelists may use particular words or phrases together repeatedly when they are searching for a latent topic online.

We identify latent topics using a skip-gram model jointly with k-means clustering.[8]

[7] Here, to reduce differences stemming from using different languages, we limit our attention to English-speaking countries only- namely, UK and US panels.

For illustration, we show examples of the sequential search terms submitted by consumers from the UK and the US panel, before and after GDPR, in Table 8. For search terms on the same topic (electronics), UK panelists submit fewer search terms before GDPR but more afterwards, while for US panelists the number of search terms used on the same topic does not increase.

We first present evidence of increased search effort by showing that there is an increase in the semantic similarity between two consecutively submitted search terms for UK panelists after GDPR in column (1) of Table 7.[9] The estimate of the interaction term indicates a significant increase in the cosine similarity between the two search terms submitted consecutively after GDPR - it implies an increase of 0.004 in the cosine similarity for UK panelists after GDPR. This suggests that after GDPR, for UK panelists the consecutive search terms showed higher degree of similarity or closer meaning over time, indicating continued search over a related o common item. This finding is suggestive of increased search efforts for these panelists after GDPR.

Second, we present evidence regarding the change in the number of keywords submitted. To identify the topic that a search term relates to, we can measure consumers' search effort on each topic by counting the total number of terms submitted on each topic and each week. For panelist i, the search effort for topic k is the logged number of terms submitted in week t, denoted by $\text{Search\_effort}_{ikt}$. In column (2) of Table 7, we see that the coefficient of the interaction term (GDPR × UK) is 0.0473 and significant at the 1% level, suggesting that after GDPR, UK panelists exerted higher search efforts in the number of keywords entered. There is an increase of 4.8% in the number of weekly keywords entered per topic, compared to their pre-GDPR average effort level. This is a second set of evidence in support of increased search friction.[10]

3.2.2 Search for Products and Services

To analyze consumer effort for product and service search, we define 'search episodes' as the sequence of products a consumer takes into consideration while searching for a particular product. Identification of a search episode from the clickstream data is not a trivial task.

[9] Semantic similarity of two search terms, is measured by the cosine similarity of their embeddings.

[10] We examine how consumer search activity changes within and between browsing sessions in Appendix C.8, Table C.11. We find that the increase in search effort comes from an increase in the number of sessions.

Table 7: Impact of GDPR on general information search effort

	Cosine similarity (1)	log(search effort) (2)
GDPR× UK	0.004***	0.0473***
	(0.001)	(0.000684)
Constant	0.459***	0.202***
	(0.000)	(0.000202)
Week FE	Yes	Yes
Panelist FE	Yes	Yes
Day of week	Yes	(not applicable)
Hour of day	Yes	(not applicable)
Topic FE	(not applicable)	Yes
No. obs	4,590,420	8,203,932
Adjusted R^2	0.14	0.19
Mean of DV	0.460	12.170

Notes: The outcome in column (1) is the cosine similarity between a search term to the previous search term submitted by the same panelist, computed using the word embeddings of the search terms. An observation is a search term. We control for panelist, week, day of the week, hour of the day fixed effects. The analysis is done at the search term level. The outcome in column (2) is the search effort, measured by number of search terms submitted under the same latent topic by a panelist in a week. We include week, panelist, and a topic fixed effects. We filled the panel so that each user-topic has the same number of weeks of appearance - we assign 0 to a topic - week if a panelist submitted zero search terms. We add 1 to the search effort and take the log. Analysis is done at user-week-topic level. Heteroscedasticity-robust standard errors clustered at user-topic level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

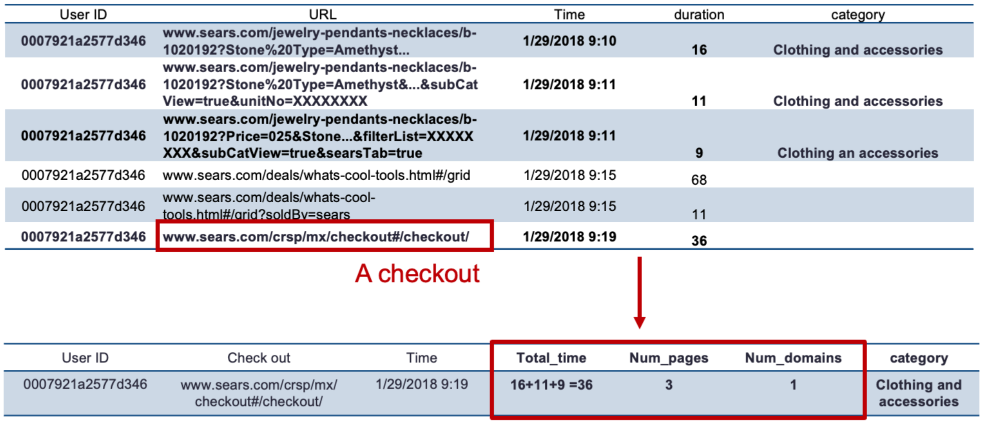
Table 8: Sample search episodes from US and UK, before and after GDPR

	UK			US	
	User 1	Time		User 2	Time
Before GDPR	samsung galaxy tab 2 310 sales	2/8/18 9:16	Before GDPR	samsung galaxy j3 luna 5 0 lte	4/28/18 22:21
	samsung galaxy tab 2 310 sale figures	2/8/18 9:16		samsung galaxy j3 luna 5 0 lte case	4/28/18 22:24
After GDPR	samsung original qi enabled afc wireless charger galaxy s9 s9	8/29/18 20:20	After GDPR	galaxy j3 luna pro sm s327vl 7 0 update	9/3/18 8:35
	samsung original qi enabled afc wireless charger galaxy s9 s9 currys	8/29/18 20:21		galaxy j3 luna pro os update	9/3/18 8:40
	samsung original qi enabled afc wireless charger galaxy s9 s9	8/31/18 9:19			
	User 3	Time		User 4	Time
Before GDPR	iphone se screen size vs google pixel 2 xl	5/5/18 13:05	Before GDPR	samsung 55 inch led 2160p smart 4k ultra hd tv best buy	2/16/18 14:45
	iphone se to google pixel 2 xl	5/11/18 10:18		samsung 55 inch led 2160p smart 4k ultra hd tv best buy	2/26/18 21:31
	iphone se to google pixel 2 xl size	5/11/18 10:19		samsung 55 inch led 2160p smart 4k ultra hd tv best settings	2/26/18 21:41
After GDPR	samsung galaxy s9 vs google pixel 2 xl	5/30/18 13:36	After GDPR	how to cast to a samsung smart tv	5/7/18 20:39
	samsung galaxy s9 advert man drop call	5/30/18 13:43		how to cast to a samsung smart tv from pc	5/7/18 20:39
	google pixel 2 xl or samsung galaxy s9	5/30/18 15:15			
	samsung galaxy s9 what s in the box	5/31/18 10:54			
	peformance test iphone se samsung galazy s9	5/31/18 12:36			

In particular, first, we must identify which browsing activities are related to products and services and which are not. Second, we must identify which consecutive browsing activities are related to the same product category and which are not. Finally, we need an indicator for whether the search came to a halt due to check out. To achieve the first step, we parse the visited URLs and search for phrases that indicate that the page contains information on a product or service. To achieve the second step, we check consecutive page visits to related product categories. Finally, for the last step, we parse the URLs for indication of checkout information, and we measure a panelist's search effort for the same type of product in a 48-hour window before a checkout page.

To identify which clicks are to product pages, and what types of products are being investigated, we utilize Google Merchant's product category taxonomy for the US and for the UK (Google Inc, 2021a,b). We parse the URLs and look for mentions of product category names, and if a name appears, we identify that page as a product browsing page under the corresponding category. Similarly, we identify checkout pages by parsing the URLs.[11] By further parsing the URLs of the checkout page (or the pages visited before), we can generally identify the category of the product a consumer checked out.[12] In total, we identified 12,381 product checkouts. Detailed summary statistics of checkout-specific search efforts and the distribution of checkouts over product categories are provided in Figure B.4 and Table B.4.

Figure 1: Illustration: Parse URLs to identify product browsing records



User ID	URL	Time	duration	category
0007921a2577d346	www.sears.com/jewelry-pendants-necklaces/b-1020192?Stone%20Type=Amethyst...	1/29/2018 9:10	16	Clothing and accessories
0007921a2577d346	www.sears.com/jewelry-pendants-necklaces/b-1020192?Stone%20Type=Amethyst&...&subCatView=true&unitNo=XXXXXXXX	1/29/2018 9:11	11	Clothing and accessories
0007921a2577d346	www.sears.com/jewelry-pendants-necklaces/b-1020192?Price=025&Stone...&filterList=XXXXXXXX&subCatView=true&searsTab=true	1/29/2018 9:11	9	Clothing an accessories
0007921a2577d346	www.sears.com/deals/whats-cool-tools.html#/grid	1/29/2018 9:15	68	
0007921a2577d346	www.sears.com/deals/whats-cool-tools.html#/grid?soldBy=sears	1/29/2018 9:15	11	
0007921a2577d346	www.sears.com/crsp/mx/checkout#/checkout/	1/29/2018 9:19	36	

User ID	Check out	Time	Total_time	Num_pages	Num_domains	category
0007921a2577d346	www.sears.com/crsp/mx/checkout#/checkout/	1/29/2018 9:19	16+11+9 =36	3	1	Clothing and accessories

Figure 1 provides an example of how we identify product pages and checkouts from the clickstream data, and how the measures of checkout-specific search efforts are created. In

[11] We look for mentions of "payment" and "checkout" in URLs.

[12] We check at most 5 pages before the checkout page to identify the product category that is purchased.

Figure 1, the URL of the visited page on top row is from sears.com and contains the words "jewelry-pendant-necklace." The consumer investigates the product for 16 seconds. The use of the phrases "necklace," "pendant necklace," and "jewelry" indicates to the researcher that the page visited contains information about a pendant necklace and thus falls under the product category "clothing and accessories". Subsequent page visits were also under the same product category on the same domain, however, presumably to different individual products, therefore, the bolded URLs indicate a search episode as they belong to the same category. For this checkout page in Figure 1, we compute the search efforts exerted by this panelist to reach the checkout: in the example, there are 3 pages clicked on, in total the product browsing took 36 seconds, and finally, one unique domain is visited under the same category, prior to the checkout.

We use three measures of search effort: total number of pages clicked, total time spent, and the number of unique domains visited in a week, as well as relating to the checked-out product category by the same panelist in the 48-hour window before the checkout. In addition, we also look at search efforts that do not end with a transaction, using the same metrics.

In Table 9, we present the results for all product browsing across sites (independent of whether it resulted in a transaction) where coefficients for $\text{GDPR}_t \times \text{UK}_i$ are 0.101 for the number of pages (column (1)), 0.262 for total time (seconds) spent on browsing (column (2)) and 0.060 number of unique domains visited (column (3)) by a panelist in a week, conditional on a product category. All estimates are significant at the 1% level. These results show that there are increases in all three aspects of product search: there is a 10.6% increase in product-browsing-related page visits, a 30.0% increase in total seconds spent on browsing in the same product category, and a 6.2% increase in the number of unique domains visited, consistent with the increase in consumer online activity (see section 3). In summary, consumers exert more effort for product search in EU regions after GDPR relative to their non-EU peers.[13] One desirable outcome of extensive search may be that, smaller brands, which typically have more limited means to reach out to consumers, may now be included in the consideration set as product search is more extensive. Put differently, e-commerce may be less concentrated

[13] As consumers begin to explore a larger number of domains and browse a higher number of pages for the same product category compared to pre-GDPR time, it is possible that they end up with a larger number of checkouts. We indeed find an increase in the consumer's number of checkouts after GDPR, and we provide details of GDPR's impact on number of checkouts in Appendix C.2.

Table 9: Impact of GDPR on product browsing

	log(no. pages) (1)	log(total time (seconds)) (2)	log(no. domains) (3)
GDPR × UK	0.101***	0.262***	0.0602***
	(0.003)	(0.007)	(0.002)
Constant	0.501***	1.297***	0.307***
	(0.001)	(0.003)	(0.001)
Week FE	Yes	Yes	Yes
Product category FE	Yes	Yes	Yes
Panelist FE	Yes	Yes	Yes
No. obs	1,264,140	1,264,140	1,264,140
Adjusted R^2	0.21	0.21	0.26
Mean of DV	3.002	171.034	0.689

Notes: The outcomes are the three measures for search efforts on a product category k made by panelist i in a week t: a panelist's number of pages clicked (column (1)), total time in seconds spent on browsing (column (2)), and number of unique domains visited under the same category (column (3)), in a week. Analysis done at week-user-product category level. We control for panelist, week, and product category fixed effects. Heteroscedasticity-robust standard errors clustered at user-category level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

or more competitive. In Section 4 we show that large sites see a greater increase relative to the small sites, therefore search frictions do not necessarily reduce concentration in the e-commerce market.

Table 10 presents estimates from Equation 4, where we focus on the search effort exerted by a panelist in the 48-hour window before a successful checkout in the same product category.[14] Conditional on a successful checkout, the number of pages, domains visited, and total time spent (in columns (1)-(3)) are lower for EU users after GDPR relative to the levels before, and the coefficients are all marginally significant. On average, compared to US panelists, UK panelists click on 1.9% fewer pages (column (1)), spend 0.21% less time on these pages (column (2)), and finally visits 1.34% fewer domains before a successful transaction (in column (3)), in a product category. Here, since the number of checkouts is much smaller compared to the overall number of site visits, we lose power in the analysis. As a result, we control for product category fixed characteristics but not for individual panelist fixed effects.

While the results in Table 9 and Table 10 are seemingly contradictory, three explanations fit this pattern, supported by additional evidence in Section 5. The first explanation is that

[14] In this analysis, unfortunately, some product categories do not have a sufficient sample size of checkouts. As a result, we dropped the categories camera and optics, baby and toddler, luggage and bags, and mature. For the details of how many checkouts are in each of the categories, see Figure B.4 in Appendix B.

Table 10: Impact of GDPR on checkout-specific search effort

	log(no. pages) (1)	log(total time (seconds)) (2)	log(no. domains) (3)
GDPR × UK	-0.099***	-0.183***	-0.034*
	(0.034)	(0.051)	(0.018)
GDPR	-0.046	-0.151	0.004
	(0.102)	(0.153)	(0.053)
UK	-0.044*	-0.123***	0.039***
	(0.025)	(0.037)	(0.013)
Constant	2.106***	5.075***	1.137***
	(0.086)	(0.129)	(0.044)
Product category FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Day of week FE	Yes	Yes	Yes
No. obs	12,284	12,284	12,284
Adjusted R^2	0.10	0.10	0.036
Mean of DV	16.456	709.325	2.913

Notes: The outcomes are the three measures for search length of a product category k made by panelist i in a week t: a panelist's number of pages clicked (column (1)), total time in seconds spent on browsing (column (2)), and number of unique domains visited under the same category (column (3)), in a week. Here, GDPR is a dummy indicating if a checkout occurred on or after the day of May 25th, 2018. An observation is a checkout, and the analysis done at checkouts level. We control for week, day of week, and product category fixed effects in all columns. Heteroscedasticity-robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

although consumers are exposed to additional alternatives through their search, as most of these search results come from inefficient searches, they do not ultimately result in a purchase. Consumers in this case may continue to buy from familiar sites, and for these sellers they need less search activity. Second, GDPR's effect may be heterogeneous across websites. If some websites are more frequent shopping destinations than others, and these sites also have different characteristics than other sites, we may anticipate GDPR to have a different impact on these sites. For instance, if bigger firms are also the ones that are more frequent shopping destinations, while the average website may be more negatively impacted by GDPR, a sample which contains these larger, more-frequently shopped domains and their pages may show, on average, a more positive GDPR impact. In Section 4, we test the impact of GDPR based on firm size, and find that larger, more frequently visited domains see less negative effects from GDPR compared to smaller, less-frequently visited domains, counter to intuitive expectation. Finally, heterogeneity among the consumers could be a factor driving the above patterns. Specifically, consumers who make more purchases online may be impacted by GDPR differently than those who purchases online less frequently. We examine whether GDPR's effects on consumers' online activity - search and browsing - are different for people with different numbers of successful checkouts before the enforcement of GDPR. Specifically, Figure C.3 shows that for panelists who have a higher number of purchases online, GDPR's effect is larger (more negative) on their search efforts while for panelists with lower number of purchases online, the decline in their search efforts are smaller.[15]

4 Heterogeneity of Effects

The increase in consumer activity online does not imply that all websites experience GDPR effects similarly. In this section, we document the heterogeneity of GDPR's effects based on domain sizes.

4.1 Effects by Domain Size

To compare small and large domains, as a measure of domain (company) size, in line with the literature, we use the number of employees of a domain. We obtain company size information from Crunchbase and Bureau van Dijk, and we link a domain to a company by matching domain names to company homepage URLs. We present the distribution of sizes in Figure B.3 in the Appendix B.4.

We employ the following specification:

$$\log(\text{traffic}_{jt}) = \gamma_0 + \gamma_1 \text{GDPR}_{jt} + \gamma_2 \text{GDPR}_{jt} \times \text{EU-penet}_j + \theta_j + \tau_t + \epsilon_{jt} \tag{5}$$

where traffic_{jt} is the unique users that site j receives in week t and EU-penet_j is a proxy measuring the relative exposure of site j to GDPR regulations. It is calculated by the share of EU panelists relative to all panelists visiting site j in the first quarter of 2018, as detailed in Table B.1 in Appendix B.4. We include week and website fixed effects to account for the time-invariant characteristics of websites and weekly general trends.

[15] We also show in Table C.2 the effect of GDPR on a panelist's browsing activity - breadth and intensity - differs by how much a panelist shops online prior to GDPR: for a panelist with more checkouts observed before GDPR, the effect of GDPR on her browsing is smaller, i.e., her search is more efficient compared to a panelist with fewer checkouts before GDPR. We relegate the details to Appendix C.3.

In this specification, GDPR's effect is measured by the coefficient γ_2, which is the percentage change in weekly traffic of a website after GDPR, compared to its pre-GDPR average traffic. Here, the GDPR date is domain-specific. Thus, identification relies on comparison of high and low EU-penetration domains, before and after GDPR compliance.

In Table 11, we present results from estimating Equation 5 on domains with different employee sizes, where employee sizes fall below and above the 90th percentile of the distribution, which corresponds to 3000 employees. [16] In column (1) the interaction is -0.045, significant at the 1% level, indicating a decline of 4.6% in weekly traffic. In column (2), the coefficient of the interaction term is 0.0003, however, not significantly different from zero. The estimates from the two columns suggest that the largest 10% of domains experience no significant change in their weekly traffic, while the rest experiences a decline. This highlights the ability of the large domains to steer clear from the negative effects of GDPR that are felt by the average company in the majority. We report similar results for different fixed effect specifications (see Appendix C.4). When we use the number of unique visitors to a website in the first quarter as an alternative measure of domain size, we find similar results (see Appendix C.5).

4.2 Heterogeneity in GDPR Effect & E-commerce Transactions

Next, we test if large and small e-commerce domains are impacted by GDPR differently. We define an e-commerce domain by checking if any pages visited on the domain contain a product category name, and identify 37,351 e-commerce sites in total. We run Equation 5 replacing the outcome variable with number of checkouts, separately for e-commerce sites above the 90th percentile and below for employee size.

In Table 12, we report estimates on subsets of e-commerce sites where employee sizes fall below and above the 90th percentile of the distribution.[17] Both columns present significant and positive estimates for two-way interaction of GDPR and EU-penetration. However, the magnitude of the estimate from column (2) is more than 10 times larger than that of column (1). These estimates imply a 1.9% increase for smaller e-commerce sites and a 13.4% increase for larger e-commerce site, where the latter is more than 6 times of the former. This result

[16] We provide results where the cutoff is the median, or 34 employees, in Appendix C.5.

[17] We provide results where the cutoff is the median of the e-commerce sites, or 76 employees, in Appendix C.5.

Table 11: GDPR's impact on website traffic by website size (employee size)

	log(weekly domain traffic)	
	Employee size below 90th percentile (3000) (1)	Employee size above 90th percentile (2)
GDPR × EU-penet	-0.0450***	0.0003
	(0.0012)	(0.0063)
GDPR	0.0233***	0.0546***
	(0.0052)	(0.0158)
Constant	0.2515***	0.5854***
	(0.0027)	(0.0084)
Domain FE	Yes	Yes
Week FE	Yes	Yes
No. obs	1,069,236	83,520
Adjusted R^2	0.83	0.91
Mean of DV	1.628	8.636

Notes: The outcome is number of unique visitors to a website in a week. The two columns present the results to the same regression specification on different subsets of the domains with different employee sizes. In all columns we control for domain and week fixed effects. Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

is consistent with our findings from Table 11: larger domains are affected by GDPR less negatively compared to smaller domains.

The findings in Table 12 also indicate that there is an increase in the total number of checkouts after GDPR. We examined the change in number of transactions made by consumers after GDPR and find that consumers have more transactions compared to pre-GDPR period. We provide the results in Appendix C.2.

4.3 Heterogeneity wrt Product Category

We next test if the estimated effect of GDPR on product and service searches is heterogeneous across product categories. We report the product category-level GDPR effects by running specification Equation 4 separately for the 21 product categories used in product search classification. Figure 2 presents the estimates of GDPR on the search efforts (number of pages browsed, total time spent, and number of unique domains visited) for each of the 21 categories. Most categories experience increase in search efforts, which are consistent with the findings in Table 9 - there are longer searches and higher search efforts in all product categories except for a few with lower number of observations (i.e., "baby toddler," "luggage

Table 12: GDPR's impact on number of checkouts at e-commerce sites

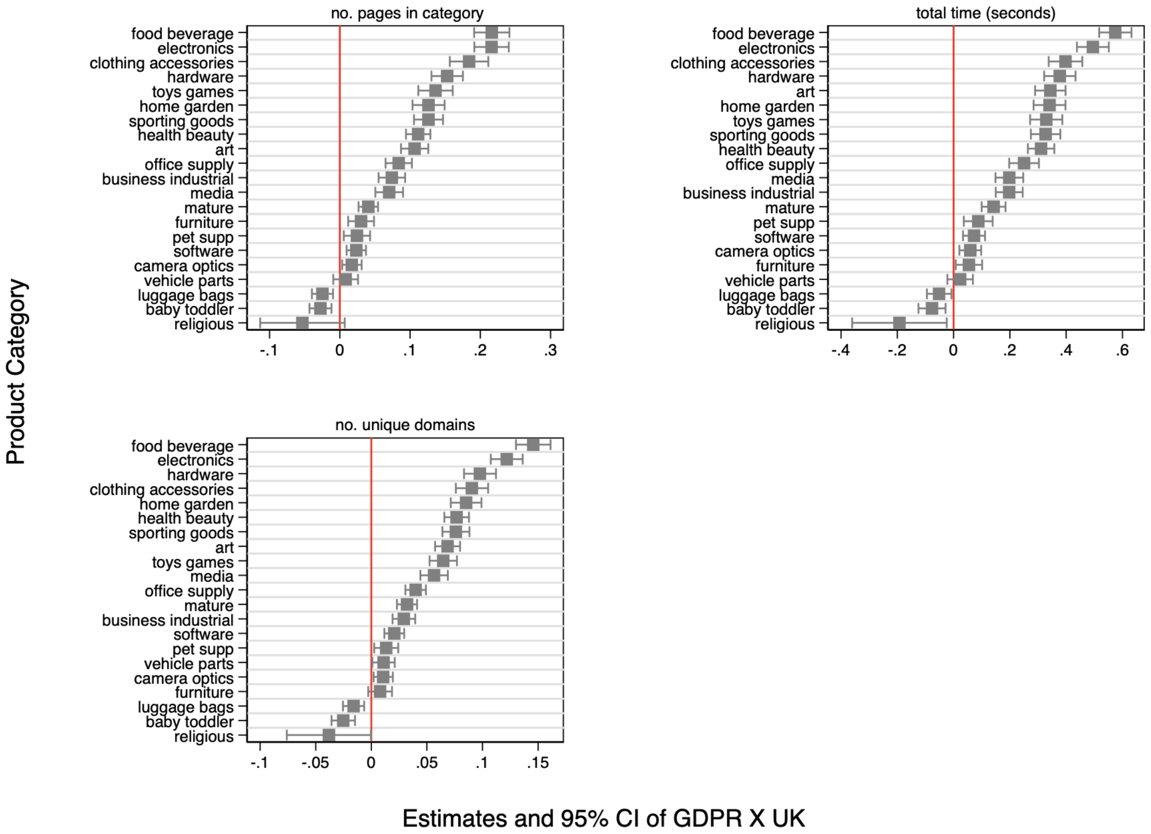
	log(weekly number of checkouts)	
	Employee size below 90th percentile (7501)	Employee size above 90th percentile
	(1)	(2)
GDPR × EU-penet	0.019**	0.126***
	(0.006)	(0.027)
GDPR	0.031**	0.074*
	(0.011)	(0.041)
Constant	0.205***	0.596***
	(0.006)	(0.022)
Domain FE	Yes	Yes
Week FE	Yes	Yes
No. obs	146,088	14,904
Adjusted R^2	0.60	0.83
Mean of DV	1.898	19.613

Notes: The outcome is number of checkouts at an e-commerce site in a week. In column (1), the sample includes domains with less than or equal to 7,501 employees which is the 90th percentile of the employee size distribution for e-commerce sites. In column (2) the sample includes domains with more than 7,501 employees. In both columns we control for domain and week fixed effects. Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

bag," and "religious"). The results show that the direction of GDPR's effect on consumer's product search is consistent across product categories. While the majority of the product categories sees increases in consumer search efforts, the magnitudes vary. The categories that experience the largest increases are "food and beverage," "electronics," and "clothing and accessories." Simultaneously, there are several categories experiencing negative GDPR effects: "luggage and bags," "baby and toddler," and "religious." These are fewer categories with fewer customers making less frequent purchases.

We offer two possible explanations for the observed heterogeneity in GDPR effects across product categories. First, the heterogeneity may stem from the prior experience consumers have with each category. For product categories consumers purchased prior to GDPR, the effect of GDPR is expected to be lower since consumers presumably have some familiarity and knowledge about their preferences, and have a more precise expectation for the value brands and products offer. Some such categories where consumers make repeat purchases include "pet supplies" and "baby and toddlers." in Appendix C.3, we provide additional evidence suggesting that consumer online browsing patterns indeed differ by pre-GDPR shopping frequency. Second, product categories may vary in their assortment size or the alternatives

Figure 2: GDPR on search efforts by product categories



Notes: We plot the estimated effects of GDPR on search efforts, or the two-way interaction from $\text{Search\_effort}_{ikt} = \alpha_1 + \alpha_2 \text{GDPR}_t \times \text{UK}_i + \eta_i + \theta_t + \nu_k + \varepsilon_{it}$: (1) number of unique domains, (2) number of pages in a category, and (3) total times (seconds) spent in a category in a week, and the 95% confidence intervals, for each of the 21 product categories. We control for panelist and week fixed effects in these regressions. The coefficients are sorted in descending order.

offered; some with a longer tail of alternatives. Such products can be expected to be more negatively impacted by search frictions, as there are more alternatives to explore (Fleder and Hosanagar, 2009; Choi and Bell, 2011). This expectation is consistent with the pattern in Figure 2 where “food and beverages,” “electronics,” “clothing and accessories”— categories with typically higher assortment volumes— see the greatest increases in search effort. Worse targeting implies worse match between consumers and products (Van den Bulte et al., 2018), and a lower likelihood of transaction, as observed in Table 12. These insights are informative for the type of websites which are more likely to be impacted by privacy regulation.

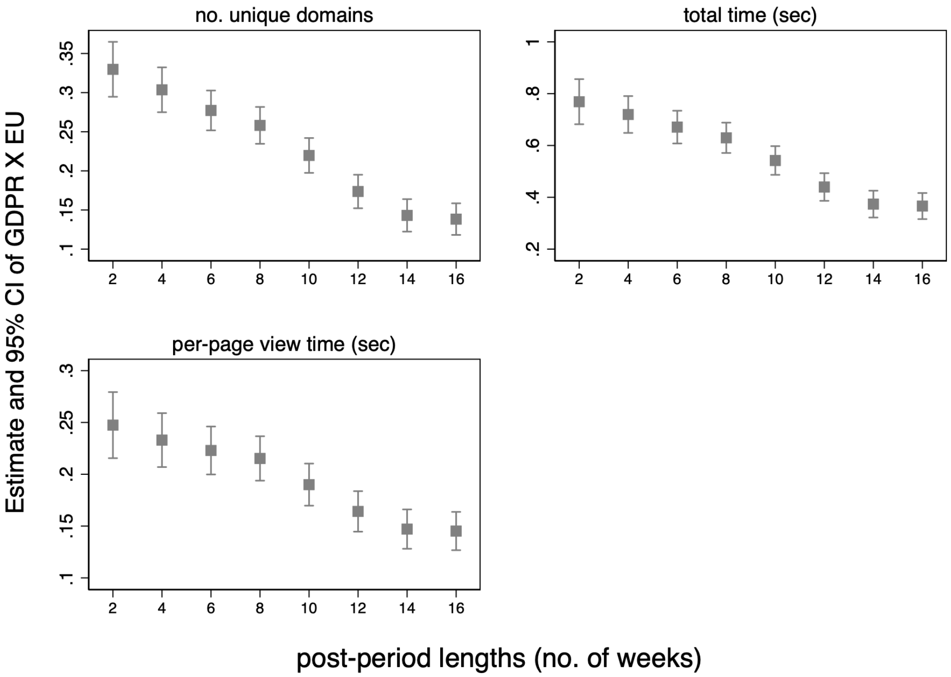
5 Robustness Checks

Window Size We first test the robustness of the magnitude of GDPR effect from Equation 1 under different window sizes. To do so, we fix the pre-period (i.e., we include all 16 weeks before the week of GDPR), and vary the length of the post-period by including up to 16 weeks after GDPR. Allowing for a longer post-period estimates the effect of GDPR on a longer horizon and thus captures the effect in the longer term, while a shorter post-period provides estimates from a short period on the outcomes. Figure 3 presents the coefficients for the interaction term $\text{GDPR}_t \times \text{EU}_i$ for the number of unique domains visited, the total time (in seconds) spent online, and finally per-page view time (in seconds). For all window sizes, the coefficients have the same sign and remain statistically significant. Second, as the window length increases, the point estimate of the interaction term declines, remaining at statistically similar levels beyond week 12. This suggests the identified GDPR effect remains positive with increasing window sizes and, while lower, remains at statistically similar magnitudes beyond week 10. These estimates for different window sizes suggest that the increase in the number of domains visited is 14.9% (from Table 4 in section 3) to 39.1% (from estimates with two-week window), 44.7% to 116% for the total time spent online, and 15.6% to 28.0% for per-page view time. However, the effects remain stable for the whole period of estimation.

Checkout-specific Searches under Alternative Window Sizes In Section 3 we assumed that all product browsing activities in the same product category that happened within the 48-hour window before a checkout page belongs to a single shopping/search episode, where the end is marked by the checkout. Figure 4 shows that our findings remain consistent if we change the window to 12-, 24-, and 36-hours.

EU-penetration Definition In Section 3 we provided GDPR's impact on domain traffic, by comparing domains with higher EU-penetration to those with lower EU-penetration. We examine if our results are robust to using alternative EU-penetration measures. Table 13 shows that the estimates of GDPR effects are consistent across columns (1) - (4), where we use four different EU-penetration measures, computed using different time periods before GDPR (first month, first two months, first quarter, and finally the first 20 weeks of 2018 - the entire period before the week of GDPR). In columns (1)-(4), the estimates of the interaction

Figure 3: Window size and GDPR effect on consumer browsing



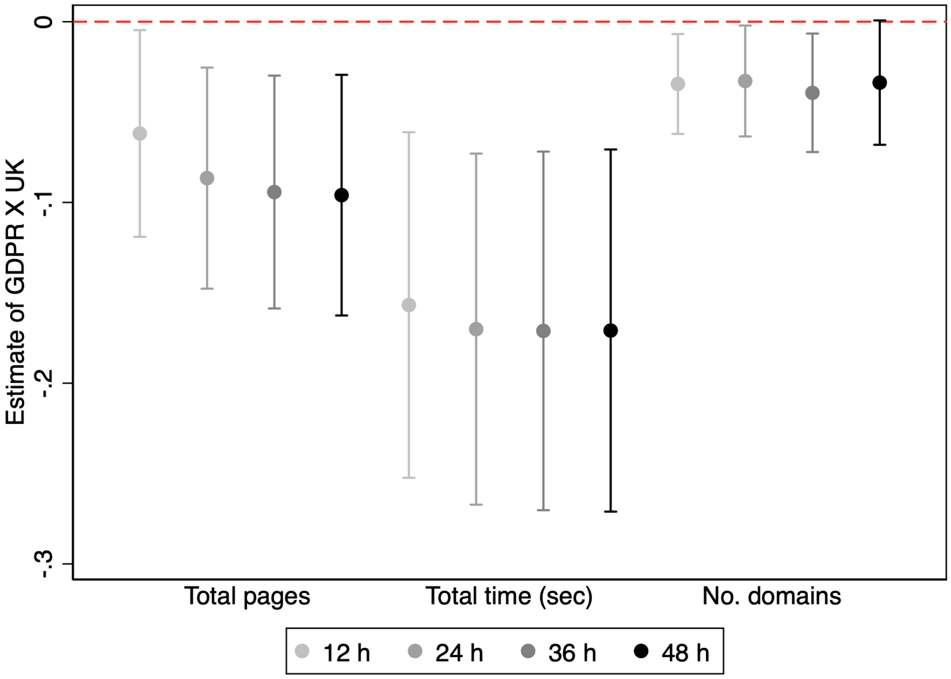
Notes: The plots are the point estimates and 95% confidence intervals of the two-way interaction term from $log(Y_{it}) = \alpha_1 + \alpha_2\text{GDPR}_t \times \text{EU}_i + \tau_t + \theta_i + \epsilon_{it}$, it measures the impact of GDPR on consumer browsing activities: number of domains visited, total time (seconds) spent online, and per-page view time (seconds). These estimates are from regressions on the sample holding the pre-period constant while varying the number of weeks in the post-period. Post-periods are the weeks after the week of GDPR. We control for panelist and week fixed effects in these regressions.

term, GDPR× EU-penet range between -0.101 and -0.050, all significant at the 1% level, and imply a 9.61% (column (1)) to 4.88% (column (4)) decline in weekly traffic.

6 Conclusion

The introduction of GDPR resulted in profound changes to the digital economy, but two that are of relevance to marketing in particular: it extended the privacy protections offered to EU residents and helped them feel safer while browsing online content (emarketer.com, 2020) and it increased the costs for firms of collecting, storing, or utilizing consumer data to track and target them via their marketing activities. The net outcome of these two opposite effects is not clear ex-ante. In this paper, we document evidence of the combined effect using a panel of consumer browsing records from four countries and investigate the changes in

Figure 4: GDPR's impact on convergent search effort vs. search windows



Notes: The plots are the point estimates and 95% confidence intervals of the two-way interaction term from $log(\text{Search\_effort}_{it}) = \alpha_1 + \alpha_2 \text{GDPR}_t \times \text{UK}_i + \tau_t + \nu_k + \epsilon_{it}$. It measures the impact of GDPR on consumer efforts exerted in a search window before a checkout page, each under four different lengths of the search window: total number of pages clicked on, total time (seconds) spent, and the number of unique domains visited in the same product category. The four estimates plotted in different grayscale ranges refer to estimates with search windows of 12, 24, 36, and 48 hours.

consumer browsing, search for information, and search for products post GDPR.

Our findings highlight that, while the EU consumers' engagement online is increasing relative to their non-EU peers after GDPR, this may not be a positive indicator overall. Further investigation into consumer search for information and products shows that, fixing topic or product category, EU consumers exert more search effort online after GDPR. The majority of the product searches do not converge to a sale, but when they do, there is a shorter, faster result for EU panelists. These findings are consistent with the explanation that higher online activity stems from a higher challenge for EU panelists to find the products and services of interest to them after GDPR. Increased costs of consumer tracking and targeting reduce the ability of firms to reach out to consumers and inform them about their products and services, such as via advertising, targeted mail, or search engine results. With GDPR, consumer search efforts also increased: they examined more pages, spent longer time

Table 13: Impact of GDPR on website weekly traffic using different EU-penetration measures

	EU-penetration in			
	(1)	(2)	(3)	(4)
	First month	First two months	First quarter	First 20 weeks
GDPR × EU-penet	-0.101***	-0.061***	-0.055***	-0.050***
	(0.001)	(0.000)	(0.000)	(0.000)
GDPR	0.007***	0.012***	0.016***	0.022***
	(0.002)	(0.002)	(0.002)	(0.002)
Constant	0.089***	0.088***	0.088***	0.088***
	(0.001)	(0.001)	(0.001)	(0.001)
Domain FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
No. obs	31,950,900	31,950,900	31,950,900	31,950,900
Adjusted R^2	0.64	0.64	0.64	0.64

Notes: The outcome is number of unique visitors to a website in a week. Columns (1) - (4) present the results to the same regression specification, but using different EU-penetration computed from first month, first two months, first quarter, and first 20 weeks of 2018. The analysis is done at website-week level and an observation is a website-week. Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

browsing a product category, and visited more alternatives while searching for a product. While consumer investigation of additional alternatives may suggest a more competitive environment online, when we investigate online transactions, we find that bigger e-commerce firms see a greater increase in the number of checkouts compared to smaller e-commerce websites.

These findings provide important insights for managers and policymakers. For marketing managers, in particular managers of e-commerce platforms, our findings suggest that they may consider intensifying their marketing efforts after GDPR. EU consumers are searching more extensively and spending more time in search after GDPR. Moreover, when they buy, they are less likely to purchase from firms they are not familiar with and from smaller e-commerce platforms. In this environment, it may be worthwhile to intensify marketing efforts.

For policy-makers, our results highlight the unintended consequences of GDPR on consumers and firms. For firms, the post-GDPR environment is anticompetitive as smaller firms see reduced consumer traffic, while for larger domains both consumer visits and consumer checkouts increase relative to the non-EU benchmark. Higher cost of compliance for smaller domains may have exacerbated the inequality between large and small domains, as evident from the differential effects of GDPR on domain traffic and e-commerce checkout volumes. For consumers, even though GDPR offers blanket privacy protections, it also introduces frictions in online browsing and search. This reduced inefficiency in search may harm consumers if it results in not being able to find the needed information or product, or results in worse search outcomes. The heterogeneity in GDPR's effects across product categories suggests that privacy regulations should take industry-specific characteristics into consideration.

While, to our knowledge, this is the first paper to demonstrate the effect of GDPR with a direct comparison of consumers in and outside the EU, our study has a number of shortcomings. In particular, for identification reasons, we are focusing on the short-term implications of GDPR. It is possible that in the long-term, magnitudes of the effects may differ, while identification is also a greater challenge. Future research can focus on this issue after identifying long term effects. Second, we document heterogeneity in GDPR effects across product categories, without taking a stance on a mechanism driving the results. Future research can complement these findings, focusing on the mechanisms behind the heterogeneity. Third, our panels focus on four countries, and naturally, some other EU countries may differ (in its market conditions and user behavior) from UK and Spain – and so the effects can be more or less severe. Future work looking into other nations can inform policymakers about the differences across EU nations regarding GDPR effects. Finally, future research may also focus on the welfare implications of GDPR by more precisely estimating these numbers.

References

Acquisti, A., Brandimarte, L., and Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347(6221):509–514.

Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8):1485–1509.

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*.

Batikas, M., Bechtold, S., Kretschmer, T., and Peukert, C. (2020). European privacy law and global markets for data. *CEPR Discussion Paper No. DP14475*.

Bronnenberg, B. J., Kim, J. B., and Mela, C. F. (2016). Zooming in on choice: How do consumers search for cameras online? *Marketing Science*, 35(5):693–712.

Choi, J. and Bell, D. R. (2011). Preference minorities and the internet. *Journal of Marketing Research*, 48(4):670–682.

Council of European Union (2014). Council regulation (EU) no 269/2014. `http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1416170084502&uri=CELEX:32014R0269`.

De los Santos, B., Hortaçsu, A., and Wildenbeest, M. R. (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review*, 102(6):2955–80.

Diamond, P. A. (1971). A model of price adjustment. *Journal of Economic Theory*, 3(2):156–168.

emarketer.com (2020). Biggest technology roadblocks to making decisions at their company according to business decision-makers worldwide. Available at `https://chart-na1.emarketer.com/240175/biggest-technology-roadblocks-making-decisions-their-company-according-business-decision-makers-worldwide-july-2020-of-respondents`.

European Commission (2019). Continuous education monitor. Available at `https://ec.europa.eu/education/sites/default/files/document-library-docs/et-monitor-report-2019-spain_en.pdf`.

Fleder, D. and Hosanagar, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5):697–712.

Goldberg, S., Johnson, G., and Shriver, S. (2021). Regulating privacy online: An economic evaluation of the GDPR. *Available at SSRN 3421731*.

Goldfarb, A. and Tucker, C. E. (2011). Privacy regulation and online advertising. *Management Science*, 57(1):57–71.

Google Inc (2021a). Google merchant center help: Product attributes. Available at `https://www.google.com/basepages/producttype/taxonomy-with-ids.en-BR.txt`.

Google Inc (2021b). Google merchant center help: Product attributes. Available at `https://www.google.com/basepages/producttype/taxonomy-with-ids.en-US.txt`.

Hern, A. and Belam, M. (2018). LA Times among US-based news sites blocking EU users due to GDPR. Available at `https://www.theguardian.com/technology/2018/may/25/gdpr-us-based-news-websites-eu-internet-users-la-times`.

Janssen, R., Kesler, R., Kummer, M., and Waldfogel, J. (2021). GDPR and the lost generation of innovative apps. *Economics of Digitization Conference, National Bureau of Economic Research 2021*.

Jia, J., Jin, G. Z., and Wagman, L. (2021). The short-run effects of the general data protection regulation on technology venture investment. *Marketing Science*, forthcoming.

Johnson, G. and Shriver, S. (2021). Privacy & market concentration: Intended & unintended consequences of the GDPR. *Available at SSRN*.

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*, 39(1):33–51.

Ke, T. T. and Sudhir, K. (2020). Privacy rights and data security: GDPR and personal data driven markets. *Available at SSRN 3643979*.

Kim, J. B., Albuquerque, P., and Bronnenberg, B. J. (2010). Online demand under limited consumer search. *Marketing Science*, 29(6):1001–1023.

Lefrere, V., Warberg, L., Cheyre, C., Marotta, V., and Acquisti, A. (2020). The impact of the GDPR on content providers. In *WEIS 2020: 20th Annual Workshop on the Economics of Information Security*.

Lin, T. (2020). Valuing intrinsic and instrumental preferences for privacy. *Available at SSRN 3406412*.

Liu, J. and Toubia, O. (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6):930–952.

OECD (2021). Average wages (indicator). doi: 10.1787/cc3e1387-en (Accessed on 06 August 2021).

Peukert, C., Bechtold, S., Batikas, M., and Kretschmer, T. (2020). European privacy law and global markets for data. *Available at SSRN 3560392*.

Seiler, S. and Pinna, F. (2017). Estimating search benefits from path-tracking data: Measurement and determinants. *Marketing Science*, 36(4):565–589.

Sorensen, A. T. (2000). Equilibrium price dispersion in retail markets for prescription drugs. *Journal of Political Economy*, 108(4):833–850.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3):213–225.

The National Statistics Institute (2017). Continuous household survey press release. Available at `https://www.ine.es/en/prensa/ech_2017_en.pdf`.

Timoshenko, A. and Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1):1–20.

Tsai, J. Y., Egelman, S., Cranor, L., and Acquisti, A. (2011). The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research*, 22(2):254–268.

Van den Bulte, C., Bayer, E., Skiera, B., and Schmitt, P. (2018). How customer referral programs turn social capital into economic capital. *Journal of Marketing Research*, 55(1):132–146.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1445–1456.

Yavorsky, D., Honka, E., and Chen, K. (2021). Consumer search in the us auto industry: The role of dealership visits. *Quantitative Marketing and Economics*, 19(1):1–52.

Appendices

(For online publication only)

A A Simple Model of Consumer Search

We illustrate the ambiguity in the effects of GDPR using a stylized model. Consider a mass of consumers, 1, each of whom is interested in visiting a set of websites. We assume that the probability of consumer i visiting website j, $P(\text{visit}_{ij})$, depends on two elements: first, the probability of website j being able to reach consumer i, $P(j \text{ reach out to } i)$, such as on search engines or via marketing activities like targeted emails and advertising; and second, the probability that consumer i visits website j upon seeing information about website j ($P(i \text{ clicks}) \; j)$. When the two events are independent, the probability that consumer i visits website j can be expressed as $P(\text{visit}_{ij}) = P(j \text{ reach out to } i)P(i \text{ clicks } j|\; j \text{ reach out to } i)$.

We model two outcomes of GDPR: enhanced protection of consumer privacy and reduced ability for firms to track consumers. Let consumer i incur a cost c_{ij} upon visiting site j due to revealing personal information such as browsing history, gender, location, etc., to the firm. GDPR lowers this cost of privacy. Specifically, let $c_{ij} = c_H$ before GDPR and c_L after GDPR, where $c_H \geq c_L$. Since tracking consumers also allows firms to target consumers more easily, let the probability the website j reaches consumer i be:

$$P(j \text{ reach out to } i) = \rho c_{ij}, \tag{6}$$

where $\rho > 0$ characterizes the technology the site uses to track the consumer, and $\rho c_H > \rho c_L$. The probability of reaching a consumer is increasing in the degree of violation of consumer's privacy. After GDPR, firms find it harder to reach consumers: $P(j \text{ reach out to } i) = \rho c_L$.

To obtain the second component of probability of a site visit, assume that consumer i's utility of visiting a website j is given by:

$$u_{ij} = v_{ij} - c_{ij}, \tag{7}$$

where her reservation utility v_{ij} is drawn from a uniform distribution on the interval [0,1]. The probability of a consumer visiting site j, conditional on a firm reaching out to consumer i, is:

$$P(u_{ij} > 0) = \begin{cases} 1 - c_H & \text{if } c_{ij} = c_H, \\ 1 - c_L & \text{if } c_{ij} = c_L. \end{cases} \tag{8}$$

Thus the unconditional probability of consumer i visiting website j is:

$$P(i \text{ visits } j) = \begin{cases} \rho c_H(1 - c_H) & \text{if } c_{ij} = c_H, \\ \rho c_L(1 - c_L) & \text{if } c_{ij} = c_L. \end{cases} \tag{9}$$

If consumers visit more websites after GDPR, it implies that $\rho c_H(1 - c_H) > \rho c_L(1 - c_L)$ must hold. However, whether this inequality holds is not clear ex-ante: when both c_H and c_L are less than 1/2, $c_H > c_L$ implies more visits to site j after GDPR, but when both parameters are greater than 1/2, the result reverses - consumers visit fewer websites after GDPR. It is ambiguous which direction GDPR's effect will go, ex-ante. These two opposing forces affect consumer's online activities observed in our panel data sets, such activities include: the breadth of consumers' visits, i.e., how many domains a consumer visits in a fixed period of time; the engagement of a consumer at a particular domain such as time of stay and number of pages viewed. If, after GDPR, it becomes more challenging for a consumer to visit websites because of information friction, he or she may visit a smaller set of websites, or exert more effort in search.

This simple model allows us to set the following empirically testable hypothesis:

H_{1a}**:** (Privacy Benefit to Consumer) With higher privacy protection resulting from GDPR, consumers are likely to explore more websites or have a higher level of engagement while visiting a website, on average.

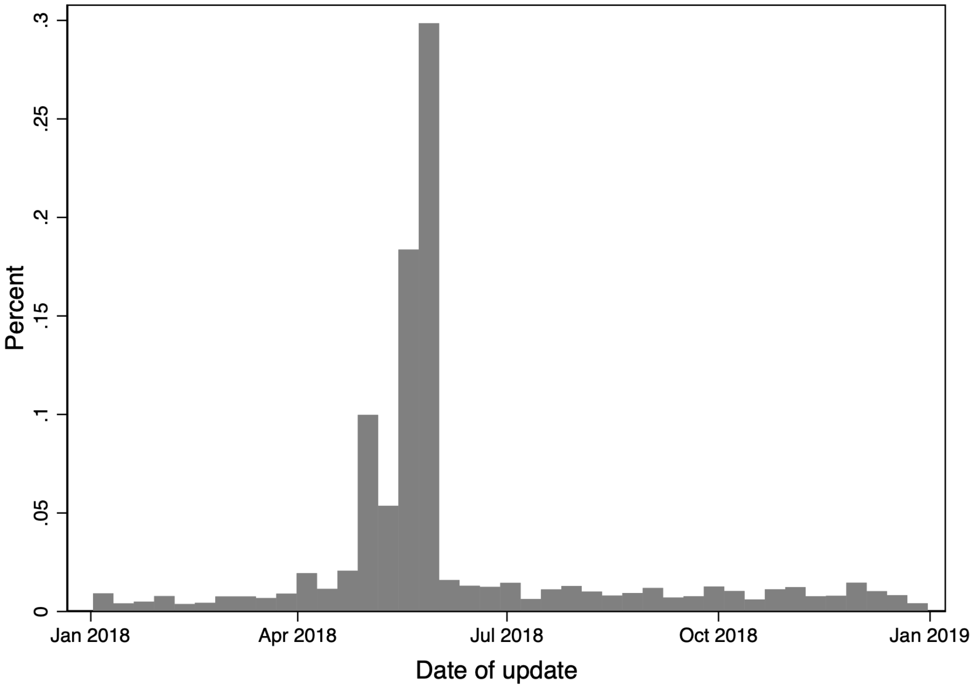
H_{1b}**:** (Friction in Search) Due to the increased cost of information collection and reduced data sharing by an average consumer, firms face friction to target consumers and consumers face friction in accessing the products and information they search.

B Data Appendix

B.1 GDPR Policy Updates

Figure B.1 demonstrates the distribution of policy update times for the websites visited in the panel.

Figure B.1: Dates of privacy policy updates



B.2 EU User Penetration and Website Traffic Data

We create a continuous measure of "EU penetration," which is a proxy for how exposed a website is to GDPR. Specifically, we calculate the proportion of traffic a domain attracts from the EU relative to all traffic in the first quarter of 2018.[18] We focus on the first quarter of the year to avoid simultaneous changes happening after GDPR goes in effect, such as some websites temporarily refusing to serve EU users (Hern and Belam, 2018). Table B.1 gives a summary of EU-penetration for all domains in our data, as well as for domains with only positive traffic data in the first quarter of 2018. Table B.2 shows the averages and standard deviations for EU-penetration for

[18] There is a long tail of 515,741 domains which receives no visits from the panel in the first quarter of 2018. We set the value of EU penetration to zero for these sites.

the first month, the first two months, the first quarter, and the first 20 weeks of 2018. These EU-penetrations from different time periods are used in Section 5 (see Table 13).

Table B.1: Summary of EU-penetration

	EU-penetration	
Sample:	First quarter traffic > 0	All domains
Mean	0.45	0.187
Std. dev.	(0.48)	(0.382)
No. obs.	371,784	887,525

Notes: EU-penetration is computed by dividing the total traffic (sum of weekly traffic) from the EU region over total traffic of a domain in first quarter of year 2018. If a domain has zero traffic during that period, we assign zero to its EU-penetration. The first column summarizes the EU-penetration of domains with non-zero traffic in the first quarter; the second column summarizes the EU-penetration of all domains.

B.3 GDPR Awareness

To provide more details of our search term data, and to show that at least some consumers in our English-speaking panels are aware of the GDPR policy, we show that there is increased number of search queries containing the word "GDPR" or "gdpr" after the official date, May 25th, 2018 as shown in Figure B.2. UK panelists search more about GDPR than their US counterparts. This also suggests that the search data reflects events that draw consumers' attention.

B.4 Firm Size Data

Data on firm size comes from Bureau van Dijk and Crunchbase databases. Bureau van Dijk reports the number of employees. Crunchbase reports employee number in intervals of 1- 10, 11 - 50, 51 - 100, 101 - 250, 251 - 500, 501 - 1,000, 1001 - 5000, 5001 - 10,000, and greater than 10,000. We take the midpoint of any interval as firm size. When a domain's parent company information is listed in only one database, we use that record. When it is available in both databases, we use the larger size of the two records because a record with smaller size is more likely to be an indicator of regional employee size, rather than the size of the whole firm. Global employee size is more likely to capture the combined resources a firm has to dedicate to GDPR changes.

Table B.2: Summary of different EU-penetration measures

	EU-penetration computed from:			
	First month	First two months	First quarter	First 20 weeks
Mean	0.026	0.114	0.187	0.307
Std. dev.	(0.157)	(0.312)	(0.382)	(0.450)
No. obs.	887,525	887,525	887,525	887,525

Notes: Summary statistics are computed from N=887,525 domains. EU-penetration is computed by dividing the total traffic (sum of weekly traffic) from the EU region over total traffic of a domain in the given period: first month, first two months, first quarter and finally first 20 weeks of year 2018. The first 20 weeks are the entire time periods before the official date of GDPR (which was in the 21st week of 2018). If a domain has zero traffic during that period, we assign zero to its EU-penetration.

Table B.3: Summary statistics of browsing activities for products and services

	Before GDPR		After GDPR	
	US	UK	US	UK
total no. pages	3.27	2.65	3.07	3.00
	(14.82)	(12.36)	(16.86)	(12.71)
total time	207.05	130.08	192.86	150.66
	(2001.31)	(1471.80)	(1940.82)	(1629.01)
no. unique domains	0.76	0.62	0.69	0.68
	(1.76)	(1.60)	(1.60)	(1.62)
No. obs	306,459	290,496	342,513	324,672

Notes: Summary statistics of search efforts, within a product category, regardless of whether there is a checkout afterwards or not.

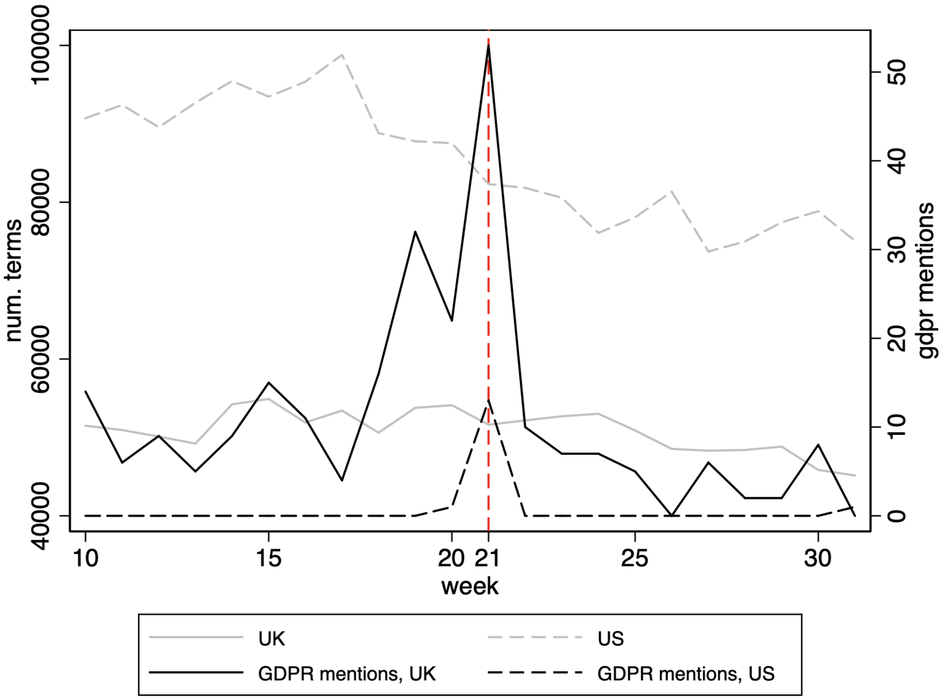
B.5 Product Search Efforts Data

In Table B.3 we present summaries of product browsing lengths, the summary statistics are broken down by the country of the panelist and by before and after GDPR. In Table B.4 we present summary statistics of checkout-specific search efforts exerted by panelists from the two countries, before and after GDPR. Just by comparing the means, checkouts made by U.S. panelists follow longer searches using all three measures, as the U.S. panelists may face a larger set of sellers/larger number of products.

B.5.1 Sample Search Terms by Clusters

In this section, we present the top three phrases and words appeared in the largest clusters identified from the search term data. To find the phrases that best represent a cluster, we look for words and phrases that appear frequently in that cluster, and less frequently in other clusters, and compute term-frequency inverse document frequency (tf-idf) score for bigrams in a cluster. We first convert

Figure B.2: Consumer searches for GDPR over time



each search term into bigrams, and we rank the phrases (bi-grams) in a cluster based on tf-idf scores. This way, we obtain phrases that are representative of a cluster. In Table B.5 we present the phrases from the 10 largest clusters identified in our data.

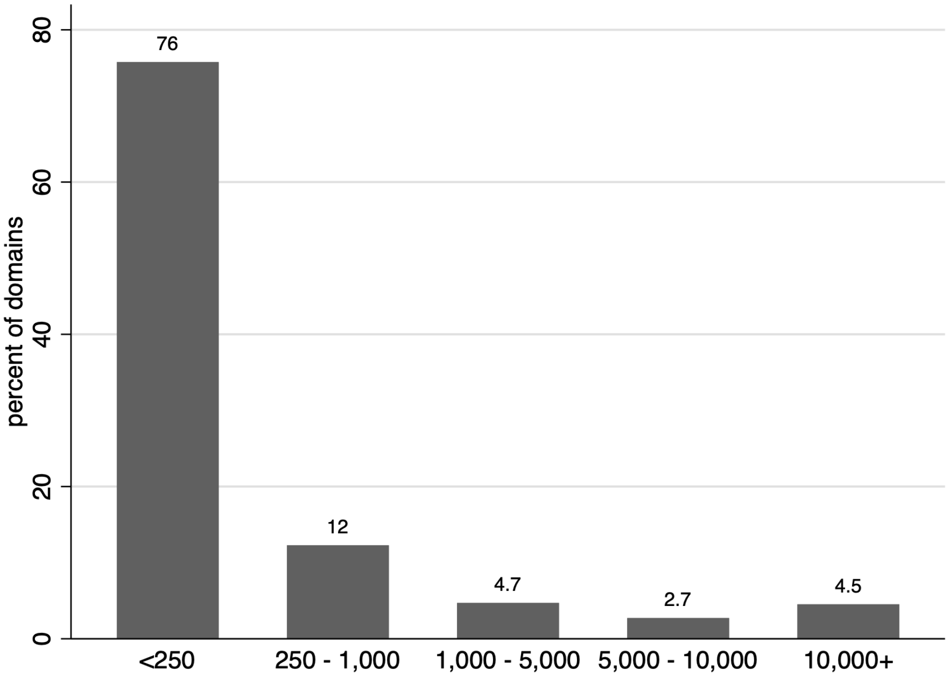
B.5.2 Checkout and Product Browsing Categories

In Figure B.4 we plot number of checkout in each of the product categories. The three categories that have the most checkouts are clothing, food, and “home and garden”. The total number of checkouts distribution to some extent speaks to whether a product category is associated with repeated purchases or random purchases, as some categories receive more checkouts than others. This heterogeneity in the purchase patterns motivates us to control for product category fixed effects in examining consumer search efforts before and after GDPR (Table 10, Section 3.2.2.).

B.6 Panelists Demographics

In this section we present summary of demographics (age, gender, education, income, and family size) of the panelists in our sample, broken down by regions. Summary statistics are in Table B.6. While we have the exact age and household size data, income and years of education are in brackets. So we estimate monthly income in US dollars by taking the midpoint of the income bracket given

Figure B.3: Employee size distribution of domains



Notes: The height of the bars indicate the percentage of firms that fall into the corresponding employee size interval. Each observation is a domain. In total there are 32,021 firms with employee information.

for a panelist. If annual income is provided, we divide the annual income by 12; if weekly income is provided, we multiply the weekly income by 4.28 (divide by 7, then multiply by 30). Finally, we convert income to local currency into US dollars, using the average conversion rate in 2018. We exclude panelists who did not report their income. For panelists who are in the highest bracket (where only the lower bound is given), we use the lower bound of their income. As a result, the income estimates may be lower than the actual income average of the panelists. For education, we create a binary variable: whether a panelist has had some college education, and we report the rate of college education among the panelists.

On average, panelists residing in the EU region are slightly older than their non-EU counterparts, are less likely to have college education, earn higher monthly income, and have smaller family sizes. Brazil's average income level is much lower than the other three developed countries. Note that the Netquest sample may not be directly comparable to the sample surveyed by the census, as the Netquest sample contains mostly adults who have access to the internet and should be on average older than the population mean.

We also present mean values from population surveys and census data (The National Statistics

Table B.4: Summary statistics of browsing activities prior to checkouts

	before GDPR		after GDPR	
	UK	US	UK	US
total no. pages	14.94	16.77	14.61	19.46
	(21.99)	(25.47)	(24.42)	(43.10)
total time (sec)	592.61	693.36	534.20	1020.12
	(1534.96)	(1642.03)	(1434.80)	(3241.72)
no. unique domains	3.10	2.86	2.94	2.78
	(3.01)	(2.54)	(3.06)	(2.34)
No. obs	2715	3339	3285	3042

Notes: The summary statistics are for variables measuring the search lengths prior to a checkout, and each observation is a checkout.

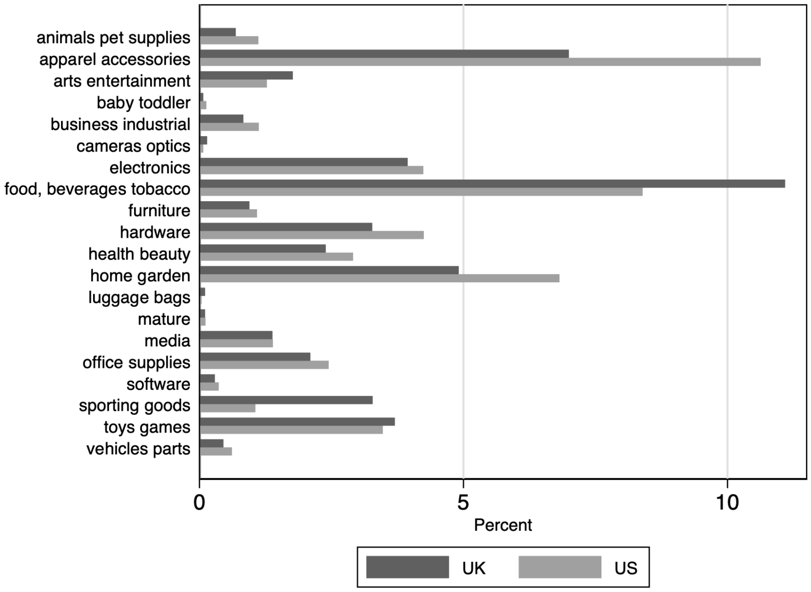
Institute, 2017; European Commission, 2019; OECD, 2021) of each country. Comparing to the census data, panelists from Brazil have higher income than the national average, while all the other panels have lower average income. All four panels show a high variance in panelists' income.

Table B.5: Words and phrases from the 10 largest clusters

Top words and phrases from UK panel			Cluster size	Cluster avg. distance
fan fiction	crazy rich asian	charles dickens	177,461	1.29418024
thirst quench	stress cause	cause headache	81,047	3.13457503
the weather outlook	temperature november	bbc weather	77,856	2.67864754
chronic fatigue syndrom	blur vision	prescribe inform	77,446	2.37114552
shoulder bag	jacket men	maxi dress	74,434	2.9736875
digital asset	paid product	bond prize	64,056	3.38855482
gingerbread latte	calori size	movie nacho	63,177	3.58617171
green lentil	cook roast	roast potato	41,531	2.83009399
road closure	waterloo road	church street	41,117	3.6347516
windows update	windows xp	fix connection	37,420	3.65105846
taylor knew trouble	music song	call music	36,531	3.18592078
seaon episode	lucifer season	handmaid tale	35,733	3.42925308
Top words and phrases from US panel			Cluster size	Cluster avg. distance
youngest billionaire	hillary duff	hayao miyazaki	137,601	3.04298786
cage sing	love lyric	sell album	112,589	3.91244585
oscar film	cruella de vil	war story	111,784	4.22841487
room decor	hand dryer	bathroom accessory	99,942	3.87970425
christmas brew	country shop	table element	94,506	3.44844328
creek michigan	island texas	diamond lake	90,794	3.94811388
prince louie	kadashian worth	meghan fall	75,599	4.44913319
new district	township michigan	house sale	75,463	4.44905809
justice department	court ban	security administrition	72,176	4.34384987
indiana realtor	louie cemeteri	obituari sunbury	71,399	3.61727885
coachella valley	dance winner	festival location	61,877	3.53670625
pie recipe	coconut cream	butter cookie	61,693	3.88756261

Notes: This table presents the top words and phrases from the 10 largest clusters identified from the data. We rank the phrases (bi-grams) in each of the clusters based on their term-frequency inverse-document-frequency scores. This way, we obtain phrases that are unique (appear frequently in its cluster, rarely in other clusters) to each of the clusters. On the right, we also present cluster sizes, measured by number of search terms in a cluster, and the average distance of a search term to the centroid of a cluster.

Figure B.4: Number of checkouts by product category



Notes: The height of the bars are the percentage of checkouts in a corresponding category. In total there are 12,381 identified checkouts.

Table B.6: Summary statistics: Panelists demographics, by EU

	US			UK			Brazil			Spain		
	Netquest	Census	Age $\geq$ 16	Netquest	Census	Age $\geq$ 16	Netquest	Census	Age $\geq$ 16	Netquest	Census	Age $\geq$ 16
Age	43.88	38.4	48.94	44.88	39.4	46.92	39.8	31.4	39.25	40.96	43.0	49.3
	(14.48)			(14.43)			(10.35)			(12.52)		
Higher education	0.70	0.88		0.45	0.61		0.48	0.37		0.50	0.42	
	(0.47)			(0.50)			(0.50)			(0.50)		
Gender (1=female)	0.60	0.51		0.59	0.51		0.49	0.51		0.48	0.50	
	(0.49)			(0.49)			(0.50)			(0.50)		
Monthly income (USD)	4,417.63	5,441		3,344.72	3,935		1,176.26	655.07		2,810.96	3,252.75	
	(3,607.87)			(2,410.08)			(668.65)			(1,707.07)		
Household size	3.01	2.53		2.76	2.39		3.33	3.30		2.97	2.49	
	(1.62)			(1.39)			(1.35)			(1.13)		

Notes: Each panel (US, UK, Spain, Brazil) includes 1,500 panelists. Monthly income is converted to US dollars using 2018's annual average conversion rate. Higher education refers to the ratio of panelists who have enrolled in college or equivalent education programs. For Spain, we report the median age instead of average age. We report conditional mean of the demographic variables for population greater than 16 years old in "Age $\geq$ 16" column if available.

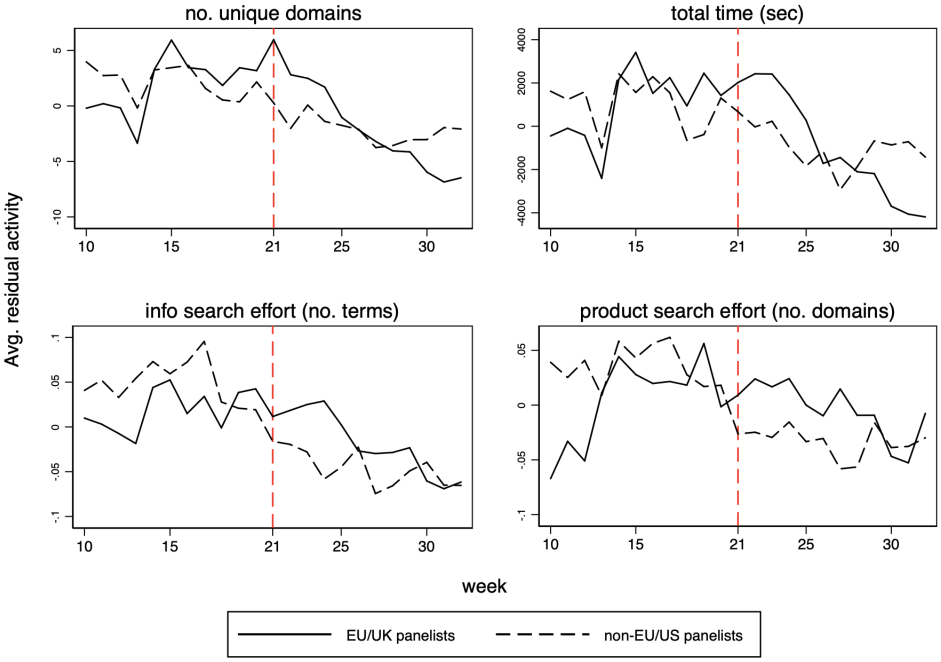
C Additional Analyses

In this section, we provide the results of additional analysis conducted as robustness checks and as supplements to our findings in the paper.

C.1 Pre-trend test

We first illustrate the parallel pretrends in Figure C.1 by plotting the weekly average residual activities (browsing, information search, and product search efforts) over weeks. The residual browsing activities (number of unique domains and total time) are obtained by regressing the outcomes on individual fixed effects. The residual search efforts are obtained by regressing outcomes on individual and topic (for information search) as well as individual and product category (for product search) fixed effects. Then, the residual activities are averaged within EU or non-EU region, for each week. The red dash line corresponds to the week of the official GDPR enforcement.

Figure C.1: Online activity over time by region



We formally verify the parallel pre-trends assumption with the specification below:

$$log(Y_{it}) = \alpha_0 + \alpha_k \sum_{k=1}^{5} GDPR_{t-k} \times EU_i + \beta_k \sum_{k=0}^{5} GDPR_{t+k} \times EU_i + \theta_i + week_t + \epsilon_{it} \qquad (10)$$

where the outcomes, Y_{it}, are the total number of unique domains visited. $GDPR_{t-k}$ is an indicator, equalling to 1 if week t is after k weeks before GDPR. For example, when k equals 2, $GDPR_{t-2}$ equals to 0 for all weeks prior to the 19th week and equals to 1 for week 19 and onwards, as GDPR is in the 21st week of 2018. $GDPR_{t+k}$'s are defined similarly.

We plot the coefficients, α_1 to α_5 and γ_1 to γ_5, together with the estimate for $GDPR_t \times EU_i$ (i.e., γ_0) in Figure C.2. These coefficients measure the difference in the outcomes for EU panelists and for non-EU panelists between two consecutive weeks, $t-k$ and $t-k-1$. Such differences become significantly different from zero if the two panels experience changes that are significantly different. In Figure C.2, for the weeks prior to the week of May 25th (i.e., the week of GDPR), we do not observe significant differences between EU and non-EU panelists in the total number of domains browsed per week, while in the week of GDPR there is a significant increase in EU user's total number of domains visited than non-EU panelists. This verifies the existence of parallel pretrends prior to GDPR in our treated group and in the control group and validates our hypothesis for the identification of GDPR effect.

C.2 GDPR's impact on consumer's number of checkouts

We show that consumers have a larger number of checkouts after GDPR. We estimate the following specification:

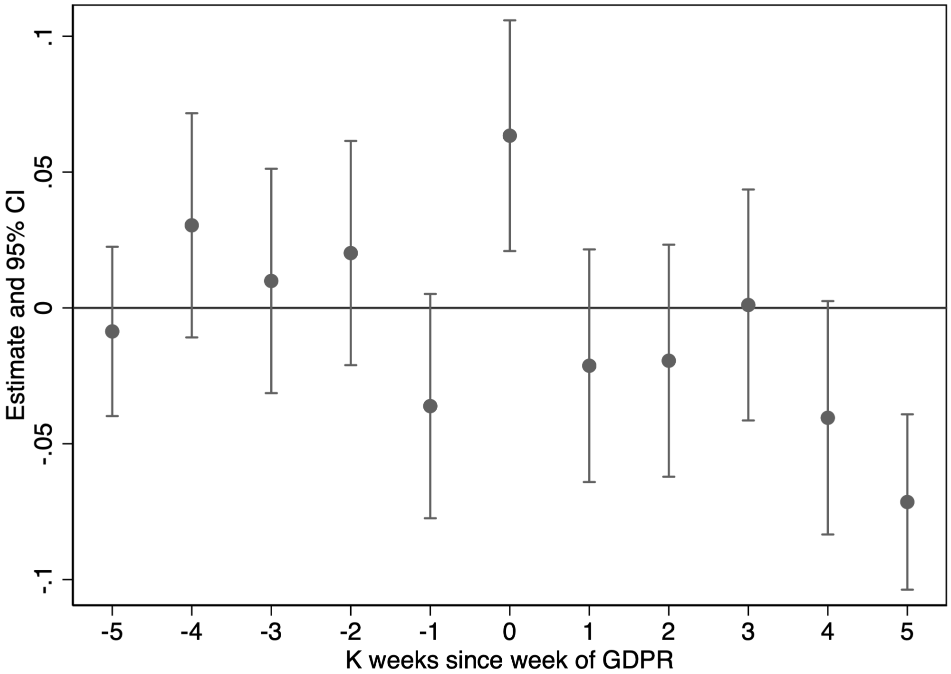
$$log(\text{no. checkouts}_{it}) = \alpha_1 + \alpha_2 GDPR_t \times \text{UK}_i + \theta_i + \tau_t + \epsilon_{it} \tag{11}$$

The dependent variable is the number of checkout pages a panelist i has in week t. If a panelist does not have any transaction pages in a week, we assign zero to that panelist-week. Checkout pages are identified the same way as in Section 4.2, i.e., by looking for words indicative of a transaction (checkout, payment) in the URLs. The identification of GDPR effect on the number of checkouts, α_2, follows our discussion in Section 4.2. In Table C.1, the interaction term estimate is significantly positive, and it implies an increase of 1.6% in number of checkouts by a consumer in a week, comparing to pre-GDPR mean.

C.3 Shopping Frequency

In Section 4.2, we find that while consumers have increased browsing activities and greater search efforts on product and services if we do not restrict to a window before checkouts. For the search efforts in a window before a checkout page, however, the search efforts exerted by the consumers declined. One possible explanation for this finding is consumer heterogeneity: consumers who have made more transactions online may have experienced different impact of GDPR than those who

Figure C.2: Testing pre-trends in logged number of unique domains visited



Notes: The coefficients plotted are the α_k's (from -5 to -1) and the γ_k's (from 0, for the week of GDPR, to 5) in the specification of $log(Outcome_{it} + 1) = \alpha_0 + \alpha_k \sum_{k=1}^{5} GDPR_{t-k} \times EU_i + \beta_k \sum_{k=0}^{5} GDPR_{t+k} \times EU_i + \theta_i + week_t + \epsilon_{it}$. The outcome is the number of unique domains visited by a panelist in a week. We control for panelist and week fixed effects. Analysis is done at user-week level, and we fill the panel by assigning zeroes to the weeks when an individual has no activity online.

have fewer or no online transactions. Consumers who have made more purchases online prior to GDPR may be easier to track and target, and could have been affected less by GDPR. We show that the effect of GDPR on consumers converging search length - product and service browsing in the 48-hour window before a checkout page - differs based on consumers' pre-GDPR shopping history. Specifically, we divide the sample of consumers with more than one checkout incidences into two subsets: those who have below 10 checkouts before GDPR (marked by white squares), and those with more than 10 checkouts before GDPR (marked by black triangles), and rerun the analysis in Table 10, Section 3.2. We plot the interaction terms of GDPR × UK in Figure C.3. For total time spent on the same product or service category, and for the total number of domains visited during the search, the estimates of GDPR effect on the two sub-samples are statistically different, and consumers who have less shopping history experience more positive change after GDPR in their search efforts. For panelists who shopped more before GDPR, the effect of GDPR

Table C.1: Effect of GDPR on consumer's number of checkouts

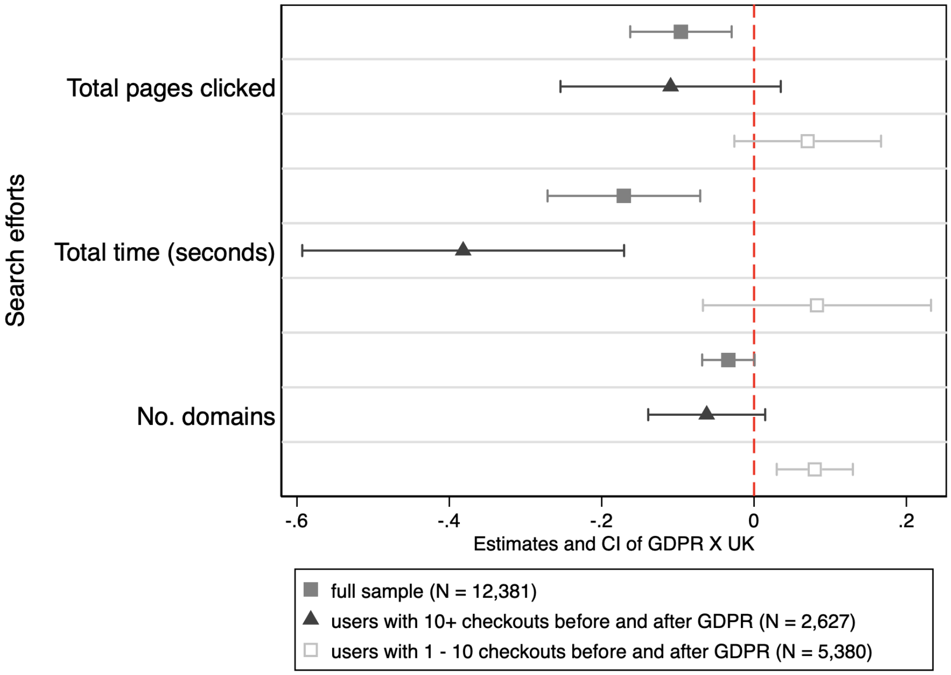
	log(no. checkouts)
GDPR × UK	0.016***
	(0.004)
Constant	0.100***
	(0.002)
Week FE	Yes
Panelist FE	Yes
No. obs	67,490
Adjusted R^2	0.19

Notes: The outcome is number of checkouts made by a consumer in a week, 0 if a panelist does not have any checkout pages in a week. Each observation is a panelist-week. We control for panelist fixed effects and week fixed effects. Heteroscedasticity-robust standard errors clustered at the panelist level in parenthesis. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

on their checkout-specific search efforts is negative: before a successful checkout, after GDPR, their search efforts declined. For the latter group of panelists who shopped less frequently, the point estimates of the effect of GDPR is positive: in the 48-hour window before a successful checkout, they exerted more efforts on searching for product-related information and visited a broader set of domains. For people who shopped more before GDPR, their search efforts decline after GDPR, which suggests that they could be more easily targeted as they could have had enough information for the firms to do so prior to the enforcement of GDPR. As a result, after GDPR, it takes them less effort to achieve a transaction. At the same time, for infrequent shoppers, they have fewer familiar firms that they had interacted with before GDPR and for them, there is a higher level of friction preventing them to reach successful purchases and their searches become more arduous.

We now test if the impact of GDPR on browsing is heterogeneous with respect to consumers' pre-GDPR numbers of purchases, or in other words, whether frequent shoppers and infrequent shoppers are affected by GDPR differently. To do so, we include a three-way interaction term of GDPR, EU, and the number of checkouts of a consumer prior to GDPR. In Table C.2, there is a marginally significant negative estimate for the three-way interactions for number of domains visited (column (2)), and significant estimates for total time and per-page time (columns (4) and (6)). These estimates show that for a consumer with a higher number of checkouts, the effect of GDPR - measured by the two-way interaction $\text{GDPR}_t \times \text{EU}_i$ - is smaller than consumers with lower number of checkouts.

Figure C.3: Effect of GDPR on consumer search efforts by pre-GDPR number of checkouts



Notes: The plotted estimates are the interaction of $\text{GDPR}_t \times \text{UK}_i$ from $\text{Search\_effort}_{ikt} = \alpha + \gamma_1 EU_i + \gamma_2 GDPR + \gamma_3 EU_i \times GDPR + \eta_i + \theta_t + \nu_k + \varepsilon_{it}$. The outcomes are three measures of search efforts prior to a successful checkout. Each observation in the data is a checkout, and the analysis is at the checkout level. In the regressions, we controlled for week and product category fixed effects. The three estimates are on the full sample (grey squares), on the checkouts made by frequent shoppers who have more than 10 checkouts before and after GDPR (black triangles), and finally less frequent shoppers with 1 - 10 checkouts before and after GDPR (white squares).

Table C.2: Change in user browsing, comparing frequent with infrequent shoppers

	log(no. unique domains)		log(total time (seconds))		log(per-page time (seconds))	
	(1)	(2)	(3)	(4)	(5)	(6)
GDPR × EU	-0.0469	-0.0621**	-0.170	-0.184***	-0.0866**	-0.0731***
× No. checkouts pre-GDPR	(0.0709)	(0.0246)	(0.113)	(0.0533)	(0.0328)	(0.0181)
GDPR × EU	0.143***	0.142***	0.390***	0.382***	0.158***	0.152***
	(0.0177)	(0.0115)	(0.0375)	(0.0286)	(0.0132)	(0.0106)
EU	-0.276***		-0.936***		-0.345***	
	(0.0131)		(0.0280)		(0.0097)	
No. checkouts pre-GDPR	0.795***	3.341***	1.244***	8.315***	0.175***	2.757***
	(0.0235)	(0.3154)	(0.0382)	(0.808)	(0.0117)	(0.272)
GDPR × No. checkouts pre-GDPR	-0.0462	-0.0457***	-0.0486	-0.0515**	0.0121	0.0077
	(0.0314)	(0.0108)	(0.0505)	(0.0216)	(0.0161)	(0.0078)
EU × No. checkouts pre-GDPR	-0.0318	-1.210**	0.264**	-2.924**	0.144***	-0.975**
	(0.0536)	(0.3822)	(0.0867)	(0.957)	(0.0248)	(0.326)
Constant	3.345***	2.871***	9.329***	7.923***	3.327***	2.810***
	(0.0055)	(0.0336)	(0.0123)	(0.0844)	(0.0045)	(0.0287)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Panelist FE		Yes		Yes		Yes
No. obs	187,092	187,092	187,092	187,092	187,092	187,092
Adjusted R^2	0.064	0.47	0.059	0.38	0.036	0.34
Mean of DV	56.396	56.396	47,620.58	47,620.58	38.721	38.721

Notes: The outcomes are three measures of weekly consumer browsing activity breadth and volume: in columns (1) and (2), the unique number of domains visited, in columns (3) and (4), total time in seconds spent online, and in columns (5) and (6), per-page average view time in seconds. Here an observation is a user-week, and the analysis is done at the user-week level. We control for week FE in columns (1), (3), and (5); we control for both week FE and panelist FE in columns (2), (4) and (6). Heteroscedasticity-robust standard errors clustered at the panelist level in the parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

C.4 GDPR Impact on Website Traffic

In Section 3.2, we have documented the increase in consumers' online browsing and search effort after GDPR. We also documented fewer domain visits for transactions which converged. We examine GDPR's effect on websites and we focus on website traffic, or unique visitors to a site in a week here. We examine both traffic that comes from desktops and visitors using mobile devices. Our analysis follows the same specification as Equation 5 in Section 4.

Table C.3 reports the estimates from this specification, and it shows that, unlike the average EU users who increased their online activity relative to their non-EU peers, websites which served a higher proportion of EU-users and therefore had higher exposure to GDPR regulations attracted, on average, less traffic compared to their peers that were less exposed. The estimate of the two-way interaction term in column (1) imply a decline of 5.35% in a website's weekly traffic after GDPR. The results together with our findings in Section 4 show that, while the EU users have increased activities online after GDPR, websites see less traffic. Two explanations may hold. First, this finding is consistent with a distributional shift in the concentration of web traffic on a smaller subset of sites, where the thinning of the long tail drives the average engagement per site down. Second, this finding would also suggest that as GDPR encourages firms to reveal use of cookies and receive their permission to collect data, consumers on the margin may be dropping out from sites, indicating a self-selection story.

C.5 Heterogeneity in GDPR Effect with Different Cutoffs for Size

In this section we present results on the heterogeneity of GDPR effect with respect to website size and e-commerce checkouts, setting the cutoff point for large and small websites at the median firm size, namely, 34 employees for all domains, and 76 employees for e-commerce sites. We also replicate the analysis in Table 11, Section 4, but use pre-GDPR traffic as an alternate measure for website size.

Table C.4 presents the effects of GDPR on websites with below and above median (34) employee sizes. Estimates in column (1) imply a decline in website traffic by 4.78% and estimates in column (2) imply a smaller decline of 3.45% for large sites.

Table C.5 presents the effects of GDPR on websites with different pre-GDPR traffic (number of unique visitors in the first quarter). The estimates from the two columns show that smaller domains whose pre-GDPR traffic fall below the 90th percentile, or the domains with less than 3 unique visitors in the first quarter, experience a 4.21% decline in weekly traffic, while larger domains above the 90th percentile of the pre-GDPR traffic distribution experience an increase of 8%. The

Table C.3: Change in website traffic

	log(weekly domain traffic)			
	(1)	(2)	(3)	(4)
GDPR× EU penet	-0.055***	-0.055***	-0.055***	-0.006***
	(0.00039)	(0.00039)	(0.00023)	(0.00036)
GDPR	0.0042***	-0.074***	0.015***	-0.0014***
	(0.00011)	(0.0026)	(0.0015)	(0.00013)
EU penet	0.14***	0.14***		
	(0.00029)	(0.00029)		
Constant	0.068***	0.03***	0.056***	0.092***
	(0.00008)	(0.00028)	(0.0002)	(0.000074)
Week		Yes	Yes	
Website FE			Yes	
Website × quarter FE				Yes
No. obs	31,981,752	31,981,752	31,981,752	31,981,752
Adjusted R^2	0.016	0.017	0.64	0.67
Mean of DV	0.33	0.33	0.33	0.33

Notes: Regression results to $\log(\text{traffic}_{jt}) = \gamma_0 + \gamma_1 \text{GDPR}_{jt} + \gamma_2 \text{GDPR}_{jt} \times \text{EU-penetration}_j + \text{week}_t + \epsilon_{jt}$ with different combinations of fixed effects. traffic_{jt} is total unique users visiting website j in week t, we take the log of this variable due to the skewness in its distribution. EU-penet_j is a continuous variable of percent of EU users website j has in the first quarter. GDPR_{jt} is a dummy indicating whether website j has already adopted a GDPR-related privacy policy by week t. We treat a website without an available updatetime to be updated on May 25th 2018. We control for week FE in all columns, website FE in column (2) and website × quarter in column (3). Each observation is a website-week record, the analysis is done at the website-week level. Heteroscedasticity-robust standard error clustered at the domain level in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

results show that the effect of GDPR on website traffic is more negative for smaller domains using both size measures.

Table C.6 presents GDPR's impact on e-commerce number of checkouts for e-commerce sites with employee sizes below and above the median (76). The estimate from column (1) imply smaller e-commerce sites - those with below median (76) employee sizes - see an marginally significant increase in number of checkouts of 1.21%. The estimates in column (2), however, indicate that large e-commerce sites experience a significant increase on 4.71%.

C.6 Missing Data

We show that our results are robust to samples with and without missing data on firm characteristics and website policy update times. Specifically, we examine if GDPR's effect is different

Table C.4: Change in website traffic (number of unique visitors) by website size

	(1) Employee size below median (34)	(2) Employee size above median
GDPR × EU-penet	-0.0490***	-0.0351***
	(0.0016)	(0.0019)
GDPR	0.0188**	0.0312***
	(0.0078)	(0.0063)
Constant	0.191***	0.361***
	(0.0041)	(0.0033)
No. obs	576,864	575,892
Adjusted R^2	0.77	0.87
Mean of DV	0.777	3.502

Notes: The outcome is number of unique visitors to a website in a week. The two columns present the results to the same regression specification on different subsets of the domains with different employee sizes. In all columns we control for domain and week fixed effects. Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

between two pairs of subsamples: (1) domains with and without privacy policy update time for GDPR compliance, and (2) domains with and without employee size information. We replicate the regression $\log(\text{traffic}_{jt}) = \gamma_0 + \gamma_1 \text{GDPR}_{jt} + \gamma_2 \text{GDPR}_{jt} \times \text{EU-penetration}_j + \theta_j + \text{week}_t + \epsilon_{jt}$ on these subsamples, and Table C.7 reports the results.

In Table C.7, in all 4 sub-samples we observe the same direction of the estimate for $\text{GDPR}_{jt} \times \text{EU-penetration}_j$: columns (2) and (3) show that the effects of GDPR, captured by the estimates to GDPR × EU-penetration, have the same sign on domains without a valid update time (so May 25th, 2018 is used as the time of GDPR), and on domains with a scraped update time (so GDPR time is domain-specific). Columns (3) and (4) show that the effects of GDPR have the same direction on domains with employee size information and on domains without employee information. The effect of GDPR on domain's traffic is robust to the missingness of website update time and employee size.

C.7 Robustness to dropping page clicks immediately after GDPR

In this section, we show that the increases in consumer browsing activities are not a result of increased time spent on website landing pages/cookie policy banners immediately after GDPR, but are a result of online information environment change. To do this, we removed the first clicks on all websites after May 25th, 2018 for all the panelists and replicated the analysis in Table 4. In Table C.10 the estimates of the interaction terms remain positive and significant, and the magnitudes are almost identical to those in Table 4. The results in Table C.10 show that the

Table C.5: Change in website traffic by website size (pre-GDPR traffic)

	(1) Pre-GDPR traffic below 90th percentile (3)	(2) Pre-GDPR traffic above 90th percentile
GDPR × EU-penet	-0.043***	0.077***
	(0.001)	(0.010)
GDPR	0.016**	-0.023
	(0.005)	(0.015)
Constant	0.148***	2.413***
	(0.003)	(0.008)
Domain FE	Yes	Yes
Week FE	Yes	Yes
Observations	1,085,400	67,356
Adjusted R^2	0.50	0.86
Mean of DV	0.278	32.121

Notes: The outcome is number of unique visitors to a website in a week. The two columns present the results to the same regression specification on different subsets of the domains with different pre-GDPR traffic - unique visitors in the first quarter. In all columns we control for domain and week fixed effects. Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

increase in consumer browsing activity is robust to removing consumers' first clicks to websites after GDPR.

We also examine whether consumer's activities on a website has changed after GDPR by replicating results in Table 3, but dropping the first clicks made after GDPR at each of the websites. The results in Table C.9 demonstrate the same sign and magnitude as those in Table 3.

C.8 Browsing and search sessions

We examine whether consumers have increased browsing time spent within a browsing session. We define a browsing session as a stream of clicks made by the same panelist, as long as no two clicks are beyond 6 hours apart. We do not find significant change with respect to number of unique domains visited per session, but the session length - in terms of total time a consumer spends in a session - has increased.

We also examine how search activities have changed within search sessions. Search sessions are defined similarly to browsing sessions but are composed of panelists' search term submissions. We look at three outcome variables: within-session average search term similarity, number of terms used in a session, total time spent on search, and number of search sessions within a week. Table C.11 present the results of GDPR's impacts on the outcomes. There are no significant changes in within-session activities, but we document an increase in the number of sessions consumers have in a week.

Table C.6: Change in number of checkouts at e-commerce sites

	(1) Employee size below median (76)	(2) Employee size above median
GDPR × EU-penet	0.012^{*}	0.046^{***}
	(0.007)	(0.010)
GDPR	0.018	0.051^{**}
	(0.014)	(0.016)
Constant	0.155^{***}	0.334^{***}
	(0.007)	(0.008)
Domain FE	Yes	Yes
Week FE	Yes	Yes
No. obs	83,088	77,904
Adjusted R^2	0.55	0.72
Mean of DV	1.145	6.091

Notes: The outcome is number of checkouts at an e-commerce site. In column (1), the sample includes domains with less than or equal to 76 employees which is the median of the employee size distribution for e-commerce sites. In column (2) the sample includes domains with more than 76 employees. In both columns we control for domain and week fixed effects. Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

Table C.7: Change in website traffic (unique visitors)

		Update time		Employee size	
	(1) All	(2) missing (use May 25th)	(3) non-missing	(4) missing	(5) non-missing
GDPR × EU-penet	-0.055^{***}	-0.055^{***}	-0.047^{***}	-0.055^{***}	-0.042^{***}
	(0.000)	(0.000)	(0.002)	(0.000)	(0.001)
GDPR	0.016^{***}		0.024^{***}	0.013^{***}	0.027^{***}
	(0.002)		(0.002)	(0.002)	(0.005)
Constant	0.088^{***}	0.093^{***}	0.273^{***}	0.082^{***}	0.275^{***}
	(0.001)	(0.000)	(0.001)	(0.001)	(0.003)
Week FE	Yes	Yes	Yes	Yes	Yes
Domain FE	Yes	Yes	Yes	Yes	Yes
No. obs	31,950,900	31,432,248	518,652	30,798,144	1,152,756
Adjusted R^2	0.64	0.63	0.78	0.60	0.85
Mean of DV	0.33	0.33	0.33	0.33	0.33

Notes: The outcome is number of unique visitors to a website in a week. Columns (1) - (5) presents the results to the same regression specification on different subsets of the domains based on whether a domain have a privacy policy update time (column (2) and (3)), or whether we are able to obtain its number of employee records (columns (4) and (5)). Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

Table C.8: Impact of GDPR on consumer browsing

	log(no. unique domains)		log(total time (seconds))		log(per-page time (seconds))	
	(1)	(2)	(3)	(4)	(5)	(6)
GDPR × EU	0.136***	0.136***	0.362***	0.362***	0.142***	0.142***
	(0.013)	(0.010)	(0.031)	(0.025)	(0.011)	(0.009)
EU	-0.300***		-0.926***		-0.327***	
	(0.010)		(0.022)		(0.008)	
Constant	3.405***	3.255***	9.536***	9.076***	3.361***	3.198***
	(0.005)	(0.004)	(0.011)	(0.009)	(0.004)	(0.003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE		Yes		Yes		Yes
Observations	187,092	187,092	187,092	187092	187,092	187,092
Adjusted R^2	0.041	0.464	0.045	0.360	0.033	0.324

Notes: The specification is $log(Y_{it}) = \alpha_1 + \alpha_2 GDPR_t \times \text{EU}_i + \alpha_3\text{EU}_i + \tau_t + \epsilon_{it}$. Each observation is a user-week record, analysis done at user-week level. Y_{it} is the number of unique domains visited by panelist i in week t, and total time panelist spends online in a week. EU_i is an indicator of whether a user is from the EU region - namely, from UK or Spain panel. $GDPR_t$ equals 0 if week t is before the week of May 25th, 2018; it equals 1 if week t is the week of May 25th, 2018 or beyond. For each of the panelist-domain pairs, we dropped the first clicks made by that panelist to that website after May 25th, 2018. Heteroscedasticity-robust standard errors clustered at the panelist level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

Table C.9: Impact of GDPR on consumer browsing activities on a domain

	No. pages	Total time (seconds)	Average time (seconds)
	(1)	(2)	(3)
GDPR × EU (user)	0.004^{***}	0.011^{***}	0.008^{***}
	(0.000)	(0.000)	(0.000)
GDPR	-0.001	0.002	0.002^{**}
	(0.001)	(0.002)	(0.001)
Constant	0.147^{***}	0.342^{***}	0.221^{***}
	(0.000)	(0.001)	(0.001)
Observations	128,337,228	128,337,228	128,337,228
Adjusted R^2	0.19	0.15	0.10

Notes: The table presents results to the specification $log(Y_{ijt}) = \gamma_0 + \gamma_1 \text{EU}_i + \gamma_2 \text{GDPR}_{jt} + \gamma_3 \text{GDPR}_{j,t} \times \text{EU}_i + \tau_t + \theta_{ij} + \epsilon_{ijt}$ where the dependent variables are (logged) number of pages viewed (columns (1)-(3)), total time (seconds) spent (columns (4)-(6)), and total number of pages clicked on a domain (columns (7)-(9)). EU_i is a dummy indicating whether panelist i is from the EU region, i.e., from the UK or Spain panels. $GDPR_{jt}$ is a dummy which equals to 1 if week t is after the week when website j updated its privacy policy in compliance with GDPR. We dropped each panelists' first click to a website after May 25th, 2018. We control for week, panelist, and domain fixed effects. We fill the panel so that for a domain-user combination, if it is missing in a week, we assign 0 to that week's outcomes. Analysis is done at domain-user-week level. In total, we have 150,789,348 observations where observations are at user-domain-week level. Heteroscedasticity-robust standard errors clustered at domain-user level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

Table C.10: Effect of GDPR on user browsing

	No. domain in a session (1)	Total time in a session (2)	No. of sessions per week (3)
GDPR × EU	-0.005	-0.014*	0.445***
	(0.004)	(0.007)	(0.009)
Constant	1.649***	4.003***	2.150***
	(0.001)	(0.002)	(0.002)
Hour of day FE	Yes	Yes	Not applicable
Day of week FE	Yes	Yes	Not applicable
Week FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Observations	1,978,649	1,978,649	187,092
Adjusted R^2	0.068	0.065	0.409

Notes: The specification is $log(Y_{it}) = \alpha_1 + \alpha_2 GDPR_t \times \text{EU}_i + \alpha_3 \text{EU}_i + \tau_t + \epsilon_{it}$. In columns (1), (2), each observation is a user-session record, analysis done at user-session level. A session is defined as a stream of clicks such that the time gap between two consecutive clicks are less than 6 hours. In columns (3), each observation is a user-week record, analysis done at user-week level. Y_{it} is (1) the number of unique domains visited by panelist i in session t, (2) total time panelist spends in a session, and (3) number of sessions in a week. EU$_i$ is an indicator of whether a user is from the EU region - namely, from UK or Spain panel. $GDPR_t$ equals 0 if session (week) t is before (the week of) May 25th, 2018; it equals 1 if session (week) t is after (the week of) May 25th, 2018 or beyond. Heteroscedasticity-robust standard errors clustered at the panelist level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

Table C.11: Effect of GDPR on search sessions

	Session level			Week level
	Avg. term similarity	No. terms	Total time	No. of sessions per week
	(1)	(2)	(3)	(4)
GDPR × UK	0.001	0.002	0.001	0.246***
	(0.001)	(0.005)	(0.009)	(0.006)
Constant	0.414***	1.411***	3.365***	0.897***
	(0.000)	(0.002)	(0.003)	(0.002)
Week FE	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes
Observations	657,058	658,211	658,211	254,700
Adjusted R^2	0.148	0.055	0.060	0.562

Notes: The specification is $log(Y_{it}) = \alpha_1 + \alpha_2 GDPR_t \times \mathrm{EU}_i + \alpha_3 \mathrm{EU}_i + \tau_t + \epsilon_{it}$. In columns (1), (2) and (3), each observation is a user-session record, analysis done at user-session level. A session is defined as a stream of clicks such that the time gap between two consecutive clicks are less than 6 hours. In columns (4), each observation is a user-week record, analysis done at user-week level. Y_{it} is (1) average cosine similarity between two consecutive search terms within a session, (2) the number of unique domains visited by panelist i in session t, (3) total time panelist spends in a session, and (4) number of sessions in a week. EU$_i$ is an indicator of whether a user is from the EU region - namely, from UK or Spain panel. $GDPR_t$ equals 0 if session (week) t is before (the week of) May 25th, 2018; it equals 1 if session (week) t is after (the week of) May 25th, 2018 or beyond. Heteroscedasticity-robust standard errors clustered at the panelist level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

D Details of Text Analysis

D.1 Search Term Pre-processing

We pre-process the search terms following several steps. First, the search keywords panel is recorded such that when a panelist is viewing the result pages after submitting a query, the same query is recorded every time the panelist browses an additional page of results. This creates several duplicate search terms but does not indicate that the panelist is starting a new search with the same keyword. With this regard we dropped, for each panelist, all redundant queries under the same domain in the same hour. Next, we dropped terms (observations from the Netquest search data) that fall into one of the following conditions: (1) only contain a domain name from the set of domains observed in the browsing data (for example, for the domain name "bestbuy.com", both queries containing only "bestbuy" and containing only "bestbuy.com" are dropped) (2) contain less than two words (3) contain more than two times of number of words than that of the 99th percentile of query length; (4) contain only numbers; (5) the words in the term on average contain more than 20 letters. Then we applied the Snowball stemmer to each word in a search query. The search queries with stemmed words are the inputs to the skip-gram model.

D.2 Details of the Skip-gram Model

In this section, we provide more details on how the skip-gram works. We begin with describing the optimization problem of the skip-gram model over word embeddings (vectors). Vector representations of words are estimated via log-likelihood function over all W words in the corpus:

$$\mathcal{LL} = \sum_{j}^{W} \{ \sum_{i:\, i,j} log(P(D_{ij} = 1|v_i, v_j)) + \sum_{i\prime \sim U(w)}^{2} log(1 - P(D_{i\prime j} = 1|v_{i\prime}, v_j))) \ \} \tag{12}$$

where the second term inside the braces $\sum_{i\prime \sim U(w)}^{2} log(1 - P(D_{i\prime j} = 1|v_{i\prime}, v_j)))$ corresponds to the two negative samples for w_j drawn from $U(w)$, the unigram distribution.

The above probability is calculated for all unique W words in a corpus. The skip-gram model has two hyperparameters, usually chosen by researchers: one is the window size, c which we detailed above, and the other one is the dimension of the hidden layer of the neural network, which is also the dimension of the word embeddings. We set the dimension of v_i's to be 200.

We prefer to use a skip-gram model because it does not require the documents observed in a corpus to be lengthy, and it expects co-occurrence of words to be at the "context" level, where a context is a sequence of $2c + 1$ or more words with c words to the left and c words to the right of the focal word, and c can be as small as just one word. Our search queries on average contain only 5.2 words, and thus the skip-gram model is more suitable for encoding query meanings. Comparing to other topic-modelling methods such as the Latent Dirichlet Allocation (LDA from here on), two features that have made LDA less suitable for our case: (1) when the documents are short, they will be converted into vectors that are sparse, as the most of the words from the entire set of words are not appearing in a search term: on average a search term in our data contains just 5.2 words while W is 119,552 for the US panel and 78,769 for the UK panel. Highly sparse vectors bias the estimates of LDA on assigning topic probabilities to that observation Yan et al. (2013); (2) LDA relies on the bag-of-word representation of texts, which does not take into consideration the sequential order of words in a document - indeed, when a document is, for example, a news article or a book chapter that contains hundreds of words, the word orders play less of an important role in deciding the latent topic, but when documents are short (search queries), it is important that we account for each word's contextual information.