Marketing Science Institute Working Paper Series 2023

Report No. 23-141

Privacy Regulations and Online Search Friction: Evidence from GDPR

Yu Zhao Pinar Yildirim Pradeep Chintagunta

Privacy Regulations and Online Search Friction: Evidence from GDPR

Yu Zhao Pinar Yildirim Pradeep Chintagunta*

This version: June 2023

Click here for the most recent version.

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 utilize recent advances in the synthetic control literature and apply the generalized synthetic control and matrix completion methods for causal inference. We document and increase in the search effort of consumers. We document decline in domain traffic through emails, and explore changes in online traffic through different marketing channels. Overall, the post-GDPR online environment 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 Associate 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, Scott Shriver, Walter Zhang, and the audiences at the 2020 Marketing Science Conference, Theory and Practice Conference, and the 2022 Strategy and Business Environment 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 the first author.

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 firms serving 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 firms collecting their data, 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 the GDPR resulted in firms being fined up to 4% of their overall revenues (e.g., BBC.com, 2019a,b; Agencia Espanola de Proteccion de Datos, 2020).

The inability to take advantage of consumer data, as reported by practitioners, results in hurdles for firms wishing to take advantage of consumer data in their marketing activities, e.g., in sending firm communications to new consumers, or targeting in advertising (Liffreing, 2018) and Joseph (2022). In turn, the inability to reach out to consumers and inform them may impact the experience consumers go through online, particularly to identify products and information that are relevant, making online search costlier, and potentially altering search outcomes.

In this study, we estimate the impact of GDPR 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., data security for 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). Enhancing consumer privacy may help consumers feel safe in their online activities, enabling them to browse and transact with more confidence. On the other hand, GDPR introduces costs for firms in collecting, storing, and utilizing consumer data, adding to informational frictions 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 hurt consumers if they have to increase their search effort in turn, or if they face worse search outcomes. It is therefore important to estimate if the benefits of GDPR due to enhanced privacy make up for the losses from increased informational friction.

We examine the net effect of GDPR using extensive online browsing and search data, with panelists from four different countries in and out of the EU region: UK, Spain, US, and Brazil. We identify the causal impact of GDPR on consumer online browsing and search, exploiting the geographical reach of GDPR. Specifically, GDPR offers protection for consumers located in the EU region (Spain and UK panelists) but not for those outside (Brazil and US panels). We first obtain baseline estimates for the effect using a difference-in-differences (DID) estimator. However, this plain vanilla DID estimator may

not yield unbiased estimates due to two concerns: the EU and the non-EU panelists may not display parallel trends prior to the issue date of GDPR, and relatedly, country-specific patterns may vary over time. For instance, the panelists in Brazil may show different seasonal consumption trends due to being in the Southern Hemisphere. To address these concerns, in addition to the results from DID, we provide estimates from panel differences (PD), generalized synthetic control (GSC), and matrix completion (MC) approaches. The first method compares a country to itself using data from pre- and post-GDPR periods, and therefore directly accounts for country-specific trends. The second method imputes counterfactuals for each treated unit (EU panelists) using data from the control group (non-EU panelists), including unit-specific intercepts interacted with time-varying coefficients. The third method, matrix completion (Athey et al., 2021), constructs counterfactual of the treated group as treating missing data problems. We will rely on the GSC method in making inference, but find that the outcomes based on the MC method often agree with those from the GSC.

To investigate if the data supports one or more of the above explanations, we then turn our attention to the change in consumers' search effort around GDPR. 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 before and after GDPR demonstrates that, the number of search terms submitted increased by 5.4% 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 9.7% more time browsing products, viewed 6.1% additional products in 3.4% more unique e-commerce sites relative to non-EU panelists. These findings are consistent with higher friction to find products.

When we look at transactions, we see that the likelihood of carrying out a transaction is not impacted negatively by GDPR. In fact, when a search resulted in a transaction, the search effort, as measured by time spent online and the number of pages viewed, was lower for the treated panelists compared to the control. These results are consistent with a set of explanations, including consumer heterogeneity and preference for buying from known alternatives, which is supported by further data. Moreover, the kind of domains where individuals are more likely to shop might have different unobserved characteristics that make them less prone to the effects of GDPR.

For policymakers, our findings imply that more strict privacy policies may result in an increase in consumer search effort online for both general information and product-related search. The implementation of the policy might have exacerbated existing differences between small and large businesses.

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. This is among the first studies that try to infer temporal search characteristics based on consumers' online browsing and keyword data, along with recent studies (e.g., Zhang et al., 2023). Given the growing interest among marketers in using machine learning methods to process consumer data (e.g., 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. In Section 2, we introduce our data sets, discuss our empirical specifications, and present results of GDPR's effect on search activity. Section 3 summarizes the heterogeneity checks we carry out in the online Appendix of this document, while Section 3.1 discusses potential mechanisms behind the observed changes. Finally, in Section 4, we conclude.

2 Data and Empirical Specification

2.1 Data

We assemble consumer online browsing panels, with firm/domain information, in addition, for our analysis. In what follows, we will provide details on each data set.

Consumer Browsing Panel: 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 include browsing panels from four different countries in and outside the EU: UK, Spain, US, and Brazil. The period over which the data are available is specific to the country, and the start dates vary between September 2017 to January 2018. In the main analysis, we will focus on the data between January 1st 2018 and September 1st, 2019, available for all four countries. This period captures the GDPR implementation date, May 25th, 2018.

The clickstream data set includes a panelist identifier, anonymized but full URL of page visits, date time of each visit to each URL, as well as the time spent at each URL.[1] We also can observe the search keywords that consumers plug into the browser to reach the domain they visit. These keywords will be used when we investigate consumer search. We will also parse the URLs to infer the products that consumers are looking at.

To eliminate the cases where a visitor may accidentally visit a site, or leave a page open in a browser without actively engaging, we drop page views shorter than 2 seconds or longer than 12 hours.[2] We concentrate on domains with at least two visitors in the first quarter of 2018. This leaves us with 95,157 domains, visited by 2,595 panelists, 685 from Spain, 576 from the UK, 565 from the US, and 769 from Brazil, respectively. For these consumers, we can also observe gender, age, level of education, household size, as well as bracketed annual income level.

Consumer Search Terms We parse consumer search queries from the URLs by looking for URL parameters containing "q" or "query", and extract the search query a panelist submits. The extracted search queries are typically phrases or short sentences, and may span a number of topics, including product-related information and more general information. We provide an example of search queries parsed from the URLs in Table A.4.

Domain Privacy Policy Change Date To check if the GDPR indeed resulted in a shift in privacy policies, 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

[1] Compared to other 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] These correspond to the 5th and the 99th percentiles of page view times, respectively.

pages.[3] We obtained update times indicating GDPR compliance for 14,551 websites. We present the website policy update times in Figure A.5 in the Appendix Section A. These scraped dates indicate that for the majority, the policy date coincides with the dates of a privacy policy change that contains terms related to the GDPR.

Domain Information 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. The summary statistics for these variables can be found in Appendix Table A.3.

2.2 Empirical Strategy

To assess the effect of GDPR on the online environments, we examine a series of consumer browsing and search activity and then firm-level outcomes. We compare weekly activities of the panelists from the EU and non-EU regions, before and after the official GDPR date. To account for unobserved heterogeneity, we include fixed effects at the panelist and week levels. This ability to track individuals over time, i.e., the panel structure of our data, is crucial to our analysis: By observing individual-level activities or by including individual fixed effects in our analysis, we are able to mitigate the concerns for alternate explanations that pertain to the differences among the panelists confounding with GDPR's effect.

As a benchmark to estimate the effects of GDPR, we start with a difference-in-differences (DID) estimator. As our panel spans the years 2018 and 2019, we are also able to conduct year-on-year analysis for EU panelists to assess the effect of the policy change via a panel differences (PD) estimator. We also adopt a generalized synthetic control and a matrix completion estimators, following the recent developments of the causal inference literature in economics.

For difference-in-differences, we employ the following specification:[4]

$$ln(Y_{ikt}) = \alpha \text{GDPR}_t \times \text{EU}_i + \text{WOY}_t + \theta_i + \epsilon_{ikt}, \tag{1}$$

where the outcomes Y_{ikt} are measures of panelist i's search efforts exerted under a topic, or a product category k by panelist i in week t. 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 which is the change in the outcome variable for the EU users after GDPR relative to non-EU users. For the DID

[3] 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.

[4] We detail the specification for PD estimator in Section B.2.

estimates from Equation 1, we include individual-specific fixed effects and week fixed effects as before.[5]

An unbiased DID estimator of the effect requires a number of assumptions. First, it assumes that the EU and non-EU panelists follow parallel trends in their online behaviors prior to the enforcement of GDPR. Second, it assumes that only the EU panelists are subject to the treatment, which may not hold if companies adopt blanket privacy policies independently of user location. Third, it abstracts away from the country-specific confounders which may corrupt the inference from the analysis.[6] To address these caveats, we will provide estimates from two additional approaches, Generalized Synthetic Control (GSC) by Xu (2017) and the Matrix Completion Method (MC) by Athey et al. (2021). Since these methods are relatively new, we will discuss their suitability to our setting before providing the results.

The GSC method works well when the treated group consists of disaggregate units (e.g., Ferraro and Simorangkir, 2020; Pattabhiramaiah et al., 2019; Bayer and Aklin, 2020) and addresses the mentioned caveats in the following ways. First, it acknowledges the presence of pretrends, based on the synthetic control method developed by Abadie et al. (2010). Second, it accounts for the individual-specific time-varying confounders. More specifically, GSC estimates the treatment effect by constructing, for each treated unit, a synthetic control counterfactual that weights the average treatment and the individual-specific time trends (Gilens et al., 2021). And, unlike Abadie et al. (2010), the method calculates standard errors and confidence intervals for ease of inference and prevents specification search via a built-in cross-validation procedure (Xu, 2017; Gilens et al., 2021).

2.3 Changes in Search Activity

We examine if the effort consumers exert to find product-specific or topical information changed after GDPR for the EU panelists. Informational frictions can alter the effort consumers put into search for various topics, e.g., news, educational materials, entertainment, as well as the effort consumers exert to find products and services that are suited to their needs. We focus on two types of searches: consumers acquiring product-related information, which we call "product-related search," and information retrieval on a topic, which we refer to as "general information search."

Ideally, to measure the search effort, the researcher would need to identify a search episode: the beginning and end time points of search, and track consumers' repeated and consecutive search input during this period for a (fixed) search outcome. The search input may include, for instance, the search terms submitted to search engines, product pages browsed before a decisive action, e.g., the domain with necessary information is visited, or a product/service transaction is carried out.

To measure the effort exerted on an information-related search (e.g., a latent topic), we infer the

[5] We do not include topic fixed effects, as topics searched tend to be panelist-specific and the panelist fixed effects absorb much of the variation across topics in our data. Data supports this thesis: the average number of topics a panelist submits to is 9 (median: 8), about 75% of the identified topics (out of 164 topics for all four panels) include less than 100 panelists.

[6] For example, Brazil experiences different seasons than the other three countries. Or, UK, Spain, Brazil have more interest in the World Cup compared to US.

topic of a search term combining the skip-gram model (Mikolov et al., 2013) together with clustering. The skip-gram model determines the vector representation of a word in a latent semantic space. A word's vector representation—called a word embedding—is chosen to optimize cosine similarity between a pair of words that appear together in the data. A query on average contains 5 words, and we take the average of these words' vectors to obtain a query's vector (Arora et al., 2017). We then perform k-means clustering on the vectorized search queries to obtain latent topics, represented by the terms belonging to the same cluster. Finally, we compute "search length" as the number of search terms submitted to the same topic (cluster), by a panelist in a given week. For details of the analysis, please visit p.A9.

To measure the effort exerted to *product-related search*, we measure the browsing intensity and breadth by counting the pages viewed, domains visited, and time spent of these pages, for pages with product names in the URLs. To identify the product names, we follow the taxonomy from Google Inc (2021) and look for the product category names in URLs. If a product name is identified, we label it as one of the 21 top-level categories. We illustrate the detail of this process in Figure B.3 in the Appendix.

Table 1: Effect of GDPR on Weekly Consumer Search Effort for Products and Information

	Product Search							
	log(No. domains)				log(Total time) (seconds)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Estimators:	DID	PD	GSC	MC	DID	PD	GSC	MC
ATT	0.0054***	0.0152***	0.1061	0.0335***	0.0270***	0.0799***	0.0930***	0.0429***
	(0.0016)	(0.0020)	(0.0723)	(0.0035)	(0.0075)	(0.0096)	(0.0315)	(0.0067)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. obs	1,219,244	927,398	1,125,456	1,125,456	1,219,244	927,398	1,125,456	1,125,456
Mean of DV	0.2192	0.2316	0.2209	0.2209	1.0751	1.1257	1.0831	1.0831
std. dev. of DV	0.4519	0.4656	0.4533	0.4532	2.1734	2.2112	2.1794	2.1794
	Product Search				Information Search			
	log(No. pages)				log(Search length)			
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Estimators:	DID	PD	GSC	MC	DID	PD	GSC	MC
ATT	0.0102***	0.0310***	0.0592***	0.0619***	-0.0225**	0.0274**	0.0529***	0.2592***
	(0.0033)	(0.0043)	(0.0133)	(0.0075)	(0.0098)	(0.0122)	(0.0184)	(0.0251)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. obs	1,219,244	927,398	1,125,456	1,125,456	141,777	111,937	131,467	131,467
Mean of DV	0.4177	0.4410	0.4208	0.4208	1.5451	1.6023	1.5458	1.5458
std. dev. of DV	0.9542	0.9800	0.9571	0.9571	0.9018	0.9454	0.9018	0.9018

Notes: The DID, PD, GSC, and MC estimates are based on 4 balanced user-week-product (topic) panels, where each observation is a user-product category-week record (columns (1) - (12)), or a user-topic-week record (columns (13) - (16)). $\log(x+1)$ transformation is applied to all outcomes.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

In Table 1, we report the results from specification (1). Overall, the results indicate a significant and positive change to search effort consumers place into finding products (columns (1)-(12)) and

general information (columns (13)-(16)). Results from product search indicate that, following GDPR, EU panelists visit up to 11% additional unique domains, spend up to 9% more time, and visit up to 6% more product pages. For general information search, all estimates except DID indicate an increase in the number of keywords inserted per search, by up to 3% additional words.

Product Purchase and Transactions. In addition to grouping search activities within a topic or a product category, we also identify the outcome of product searches by looking for checkouts in the URLs. For each of the checkouts identified, we assign the transaction to the category whose name appears within 5 pages prior to the checkout page. We retrieve product page views under that category prior to a checkout within a 48-hour window. We measure the search efforts exerted in this time window to examine, conditional on a successful transaction, whether search has become lengthier after GDPR. In Table 2, we report the DID estimators for the three measures of effort.

Table 2: Effect of GDPR on product search, 48 hours prior to checkout pages

	log(No. domains)			log(time)			log(No. pages)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GDPR × EU	0.0036	0.0096	0.0064	-0.0355	-0.0089	0.0797	-0.1286**	-0.0931*	0.0544
	(0.0231)	(0.0225)	(0.0255)	(0.0725)	(0.0695)	(0.0791)	(0.0596)	(0.0560)	(0.0628)
No. obs	8,209	8,209	8,055	8,209	8,209	8,055	8,209	8,209	8,055
adj. R-squared	0.0401	0.0870	0.3280	0.0370	0.1315	0.3330	0.0386	0.1690	0.3787
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panelist FE		Yes	Yes		Yes	Yes		Yes	Yes
Product category FE			Yes			Yes			Yes
Mean of DV	1.1507	1.1507	1.1544	5.9773	5.9773	5.9895	2.9111	2.9111	2.9266
std. dev. of DV	0.4963	0.4963	0.4977	1.5763	1.5763	1.5753	1.3039	1.3039	1.3038

Notes: Each observation is a checkout, the analysis is at the checkout level. The dependent variables are panelist product search activities in a 48-window prior to the checkout page, carried out under the same product category of the transacted item. The checkout's category is identified by product names appeared in URLs up to 5 pages prior to the checkout page. $\log(x+1)$ transformation is applied to all outcomes.
* $p < 0.1$, ** $p < 0.05$, ***

3 Firm-level Effects

We examine GDPR's effect on domains, and we focus on website traffic, or unique visitors to a site in a week. For this analysis, we use an analysis at the domain level. In particular, the identification here relies on the variation between domains in terms of the number of EU users that they have to accommodate, since the requirements by EU hold if the rights of the EU residents are violated. We approximate the exposure of an EU domain via a continuous measure of "EU penetration." Our analysis follows the specification as Equation 3:

$$\log(\text{traffic}_{jt}) = \gamma_0 + \gamma_1 \text{GDPR}_t + \gamma_2 \text{GDPR}_t \times \text{EU-penet}_j + \theta_j + \tau_t + \epsilon_{jt} \tag{2}$$

Table 3: Effect of GDPR on Consumer Product Browsing (sessions)

	No. domain (1)	No. pages (2)	Total time (3)	Checkout observed (4)
GDPR × EU	0.0025**	0.0043	0.0468**	0.0014**
	(0.0011)	(0.0050)	(0.0174)	(0.0005)
Product category × month FE	✓	✓	✓	✓
Weekday FE	✓	✓	✓	✓
GDPR × EU	-0.0001	-0.0044	0.0004	0.0012***
	(0.0012)	(0.0047)	(0.0012)	(0.0002)
Product category × month FE	✓	✓	✓	✓
Weekday FE	✓	✓	✓	✓
Country FE	✓	✓	✓	✓
GDPR × EU	-0.0039	-0.0077**	-0.0008	0.0013***
	(0.0027)	(0.0034)	(0.0035)	(0.0001)
Product category × month FE	✓	✓	✓	✓
Weekday FE	✓	✓	✓	✓
Panelist ID FE	✓	✓	✓	✓
Observations	2,564,700	2,565,904	2,564,700	2,565,904
Dependent variable mean	1.0145	1.9214	4.7130	0.01966
Dep. Var. std. dev.	0.47065	1.1600	1.9103	0.13884

Notes: Each observation is a panelist's product browsing session, a session is defined as consecutive pages viewed within the same product category, with gaps shorter than 12 hours.

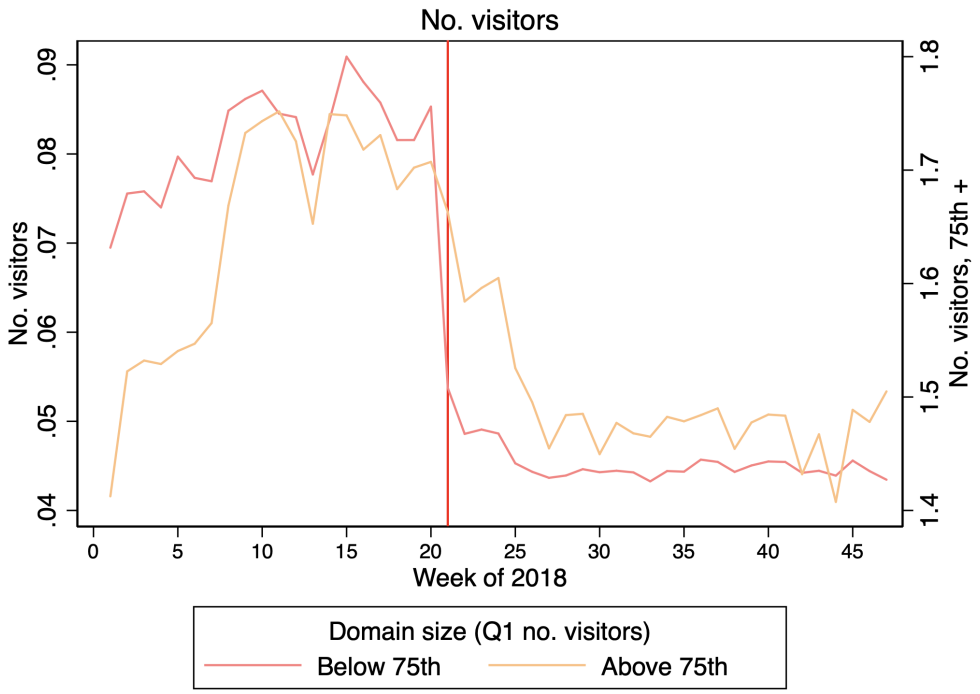
where traffic_{jt} is the unique users that site j receives during week t and EU-penet_j is a proxy that measures the relative exposure of site j to GDPR regulations. Specifically, we calculate the proportion of traffic a domain attracts from the EU relative to all traffic in the first quarter of 2018.[7] We include week and website fixed effects to account for the time-invariant characteristics of websites and weekly general trends. 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. Thus, identification relies on comparison of high and low EU-penetration domains, before and after GDPR deadline.

The effects of GDPR may vary heterogeneously across firms of different sizes and across product types. The effects may vary firstly because larger firms have more resources to allocate to compliance and consumer protections, and secondly because firms that attract larger numbers of consumers may differ in observable and unobservable ways (e.g., providing higher quality products, etc.) which may interact with consumers' desires to consent to sharing data with them. Figure 1 compares the

We report the results in Table 4, where the coefficients are estimated on samples of domains with different sizes, split 2, 6, and 103 unique visitors. We note that the first split, 2 unique visitors,

[7]Summary statistics of this measure are given in Table A.3, together with measures of domain size. We focus on the first quarter of the year to construct measures of EU exposure and domain sizes, to avoid simultaneous changes that occur after GDPR goes into effect, such as some websites temporarily refusing to serve the EU users (Hern and Belam, 2018). We also focus on domains with at least 1 unique visitors in the first quarter, as domains without any visitors from the first quarter may have a shorter pre-period compared to other domains. In addition, the very small domains may be irrelevant to the majority of the online activity.

Figure 1: Domain traffic, by pre-GDPR number of visitors



Notes: The lines represent average residual number of visitors to a site in a week, average taken over the size bins. The residuals are from regressing number of visitors on domain fixed effects. The 75th percentile of domain size (no. distinct visitors from the first quarter of 2018) is 2. The left y-axis is the average weekly traffic to smaller domains with 2 or fewer visitors in the first quarter, and the right axis the larger domains. The red vertical line marks the deadline of GDPR at May 25th, corresponding to the 21st week of 2018.

corresponds to about the 75th percentiles as the distribution of domain sizes has a long tail.

In Table 4 and Table 5, we present the results to Equation 3, controlling for domain, week fixed effects as well as domain $\times$ quarter fixed effects respectively. In both tables, the estimates show that GDPR is associated with an increase for the larger domains (columns (3) and (4)) with the largest increase to the most popular domains, though insignificant once we control for domain $\times$ quarter fixed effects: this reflects the fact that the largest domains may span a variety of industries and the domain-quarter fixed effects partially capture the heterogeneous industry dynamics over time.

In Table 5, we document different directions of change in site traffic across small and large sites. In particular, for domains with 2 or fewer visitors, or about 75% of all the domains observed in our panel, if domain's EU-penet increase by 1 unit, its traffic would decline for about 0.11% than a domain with only 54% EU visitors[8]. In Table 4, the estimates through columns (1) to (4) highlight that for the largest domains, the increase in average weekly traffic for the largest domains is beyond an order of magnitude higher than the smallest ones, and six times as large as the next largest domains with 6 - 103 visitors in the first quarter. Taken together, these results show that post-GDPR there may be

[8] A domain with EU-penet equals 0.54 would be the base as the EU-penet has a mean value of 0.53 and was normalized to have mean 1 in the regression.

Table 4: Effect of GDPR on domain traffic, by domain's size (sample: domain size ≥ 1)

Domain size:	(1) below 2	(2) 3-6	(3) 6-103	(4) above 103
GDPR × EU-penet (mean = 1)	0.0006***	0.0053***	0.0298***	0.1295***
	(0.0002)	(0.0012)	(0.0022)	(0.0133)
No. obs	13,681,155	1,569,595	1,300,090	48,548
adj. R-squared	0.1502	0.1981	0.5725	0.7681
Week FE	Yes	Yes	Yes	Yes
Domain FE	Yes	Yes	Yes	Yes
Mean of DV	0.0387	0.1630	0.6424	3.0490
std. dev. of DV	0.1644	0.3339	0.7296	1.0784

Notes: An observation is a domain-week, the analysis is at the domain-week level. The sample includes domains with at least 1 visitor in Q1, 2018. In column (1), the sample includes domain with 2 or fewer unique visitors from the first quarter, corresponding to approximately bottom 75% of all domains in our sample. EU-penet is normalized and has mean of 1. $\log(x+1)$ transformation is applied to all outcomes.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

an increase in the disparity between large and small sites / firms.

3.1 Discussion of Possible Mechanisms

The results indicate an increase in the product browsing time and the number of domains visited by the average panelist in EU relative to their non-EU peers, post-GDPR, but not higher convergence in individual's product browsing. Since the time spent per page does not change on average, the increase in overall time is likely to be driven by visits to more domains. Why do EU users spend more time and visit more domains related to search? We consider two possible explanations consistent with the observed empirical pattern.

First, it is possible that due to "enhanced privacy protections" warranted by GDPR, consumers may feel more comfortable visiting less familiar domains and spending more time online when there is more privacy. If consumers receive an additional utility from visiting each domain net of time spent, this may improve consumer welfare. Second, post-GDPR environment may also be costlier for firms to reach out to consumers, as consumers must give consent to receiving firm communications (e.g., promotional emails) and accepting cookie tracking (essential for targeted advertising). When some consumers on the margin drop out and do not give consent, ceteris paribus, a firm may expect fewer total number of visitors on average, and fewer users coming to a site via direct communication (emails). Aridor et al. (2020) reports that marginal consumers opt out of marketing communications by rejecting cookies. Due to reduced consent, the number of people who come from targeted advertising may also decline. At the same time, the change in those coming from indirect channels (e.g., mass advertising and organic search) is ambiguous. We illustrate these possible changes in Figure A.1.

We investigate how a domain's traffic share from different marketing channels has changed after

Table 5: Effect of GDPR on domain traffic, by domain's size (sample: domain size ≥ 1)

Domain size:	(1) below 2	(2) 3-6	(3) 6-103	(4) above 103
GDPR × EU-penet (mean = 1)	-0.0011***	0.0007	0.0108***	0.0175
	(0.0002)	(0.0014)	(0.0025)	(0.0128)
GDPR	-0.0232***	-0.0335***	-0.0596***	-0.0739***
	(0.0003)	(0.0018)	(0.0030)	(0.0133)
No. obs	13,681,155	1,569,595	1,300,090	48,548
adj. R-squared	0.1970	0.2590	0.6231	0.7990
Domain × quarter FE	Yes	Yes	Yes	Yes
Mean of DV	0.0387	0.1630	0.6424	3.0490
std. dev. of DV	0.1644	0.3339	0.7296	1.0784

Notes: An observation is a domain-week, the analysis is at the domain-week level. The sample includes domains with at least 1 visitor in Q1, 2018. In column (1), the sample includes domain with 2 or fewer unique visitors from the first quarter, corresponding to approximately bottom 75% of all domains in our sample. EU-penet is normalized and has mean of 1. $\log(x+1)$ transformation is applied to all outcomes.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

GDPR. To this end, we adopt a different specification:

$$\log(\text{traffic}_{jt}) = \gamma_1 \text{GDPR}_t + \gamma_2 \text{GDPR}_t \times \text{EU-penet}_j + \theta_j + \tau_t + \epsilon_{jt} \tag{3}$$

where s^c_{jt} is the share of unique visitors that site j receives through channel c during week t, where the channel is one of the following: ad-click, email, organic search, search keyword ad, and social media ads. Share of traffic from these channels, EU-penet$_j$ is a proxy that measures the relative exposure of site j to EU visitors. Specifically, we calculate the proportion of traffic a domain attracts from the EU relative to all traffic in the first quarter of 2018. Summary statistics of this measure are given in Table A.3 in the Appendix, together with measures of domain size. We focus on the first quarter of the year to construct measures of EU exposure and domain sizes, to avoid simultaneous changes that occur after GDPR goes into effect, such as some websites temporarily refusing to serve EU users (Hern and Belam, 2018). We include week and website fixed effects to account for the time-invariant characteristics of websites and weekly general trends. 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. Thus, identification relies on comparison of high and low EU-penetration domains, before and after GDPR deadline.

We examine changes in site view time for visits through the following channels: ad-click, (promotional) email, and organic search. Among the three channels, domain visits through promotional emails requires the company to obtain customer's consent of providing contact information, and we expect, as discussed in Section 3.1, the share of visits from email decrease but the average interest of these visitors increase, as consumers on the margin may opt out from the firm's marketing communications. For ads, as advertisements include both targeted ads, which relies on consumer's provision

of personal information, and broadcasting ads, which require less knowledge of the audience, traffic shares through this channel may increase or decrease. However, if the enforcement of the policy results in easier opt-out thus worsened targeting on the firm's end, we should expect to see decline in the "fit" of ads delivered to the audience, and these ad-clicks may result in shorter stay on the corresponding domains. Finally, for search, if the former two channels are losing effectiveness overall, we should expect consumers to utilize the search engine more, i.e., input more search efforts, as shown in Table C.1.

Table 6: Session length following site visits through marketing channels

	Click duration			Domain view time, next hour		
	Ad-click	Email	Search	Ad-click	Email	Search
GDPR × EU	-0.089***	-0.025***	-0.054***	-0.0656***	-0.0171	-0.0399**
	(0.005)	(0.008)	(0.010)	(0.0086)	(0.0129)	(0.0171)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Panelist FE	Yes	Yes	Yes	Yes	Yes	Yes
No. obs	1,277,432	519,134	284,622	1,375,887	541,977	304,054
Mean of DV	2.711	2.671	2.514	3.2708	2.0943	2.4541
Std. dev. of DV	1.423	1.335	1.164	2.7651	2.2288	2.2094

Notes: Each observation is a **click** to a website through one of three marketing channels: Ad-click, email, or organic search. The analysis is at the click level. The outcomes from columns (1) - (3) are logged page view time for all site pages viewed in an hour following the click, in seconds. The outcomes from columns (4) - (6) are total doamin view time in the next hour following the click. Pages shorter than 2 seconds or longer than 12 hours are dropped. GDPR indicates whether the click is on or after May 25th, 2018; EU indicates whether a panelist is from UK or Spain. $\log(x+1)$ transformation is applied to all outcomes. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$

Overall, these findings suggest changes in the overall effectiveness of the direct and indirect marketing channels, in reaching out to existing customers and in bringing new customers to domains. Specifically, we discover that organic search is more effective in introducing new customers to a site, for only the largest domains after GDPR.

4 Conclusion

The introduction of GDPR resulted in profound changes in the digital economy, but two that are relevant 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 increased the costs for firms to collect, store or utilize consumer data. 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 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 of

consumer search for information and products shows that, the 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 the EU panelists. These findings are consistent with the explanation that higher online activity stems from a higher challenge for the EU panelists to find the products and services of interest to them after GDPR. The increased costs of consumer tracking and targeting reduce the ability of firms to reach consumers and inform them about their products and services, such as through advertising, targeted mail, or search engine results. After GDPR, consumer search efforts also increased: they examined more pages, spent more time 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. The 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 focus on the short-term implications of GDPR. It is possible that, in the long term, the 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 at 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

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

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

Agencia Espanola de Proteccion de Datos (2020). Procedimiento No.: PS/00037/2020. Available at `https://www.aepd.es/es/documento/ps-00037-2020.pdf`. Accessed: 9-29-2022.

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*.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.

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

Bayer, P. and Aklin, M. (2020). The european union emissions trading system reduced co2 emissions despite low prices. *Proceedings of the National Academy of Sciences*, 117(16):8804–8812.

BBC.com (2019a). British Airways faces record £183m fine for data breach. Available at `https://www.bbc.com/news/business-48905907`. Accessed: 9-29-2022.

BBC.com (2019b). EE fined £100,000 for unlawful texts. Available at `https://www.bbc.com/news/technology-48743784`. Accessed: 9-29-2022.

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.

Cha, M., Gwon, Y., and Kung, H. (2017). Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006.

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`.

Ferraro, P. J. and Simorangkir, R. (2020). Conditional cash transfers to alleviate poverty also reduced deforestation in indonesia. *Science Advances*, 6(24):eaaz1298.

Gilens, M., Patterson, S., and Haines, P. (2021). Campaign finance regulations and public policy. *American Political Science Review*, 115(3):1074–1081.

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

Hashimoto, T. B., Alvarez-Melis, D., and Jaakkola, T. S. (2016). Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286.

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.

Joseph, S. (2022). With the future of third-party addressability on the open web hanging in the balance, the ad industry divides in two. Available at `https://digiday.com/marketing/with-the-future-of-third-party-addressability-on-the-open-web-hanging-in-the-balance-the-ad-industry-divides-in-two/`. Accessed: 2022-9-29.

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*.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27:2177–2185.

Liffreing, I. (2018). Marketers struggle to track audiences after facebook and google scale back data for gdpr. Available at`https://digiday.com/marketing/marketers-struggle-track-audiences-facebook-google-scale-back-data-gdpr/`. Accessed: 2022-9-29.

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.

Liu, L., Wang, Y., and Xu, Y. (2021). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *arXiv preprint arXiv:2107.00856*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Pattabhiramaiah, A., Sriram, S., and Manchanda, P. (2019). Paywalls: Monetizing online content. *Journal of Marketing*, 83(2):19–36.

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.

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.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

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.

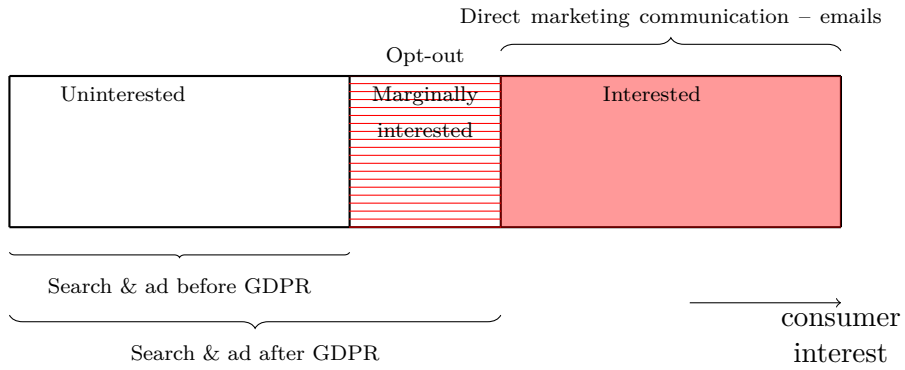
Yavorsky, D., Honka, E., and Chen, K. (2021). Consumer search in the us auto industry: The role of dealership visits. *Quantitative Marketing and Economics*, 19(1):1–52.

Zhang, X. L., Ursu, R., Honka, E., and Yao, Y. O. (2023). Product discovery and consumer search routes: Evidence from a mobile app. *Available at SSRN 4444774*.

Appendices

A Data and Descriptive Statistics

Figure A.1: Market for a domain



A.1 Definitions of Outcomes

The main and supplemental analysis in the paper and detailed in the online appendix focus on a series of outcomes at the panelist and domain levels. For panelists, we examine weekly browsing activities and on various measures of search efforts. For domains, we examine site traffic and traffic share through different channels. In Table A.1, we summarized the definitions of the outcomes we focus on in the paper, the unit of observation of different data sets, and the structure of the data sets, i.e., whether a data set is a panel or is consisted of clicks, we constructed from the clickstream (see Section 2) data.

Table A.1: Glossary of outcomes of interest

Outcome	Level	Missing value
Consumer online browsing activity: (Y_{it})		
Number of unique domains	Panelist-week	Filled with 0
Total time online	Panelist-week	Filled with 0
Per-page view time	Panelist-week	Filled with 0
No. page clicks	Panelist-week	Filled with 0
Consumer search:		
Search length, no. search terms submitted	Panelist-topic-week	Filled with 0
Search length, no. product pages for a product type	Panelist-product type-week	Filled with 0
Search length, no. domains for a product type	Panelist-product type-week	Filled with 0
Search length, total time on a product type	Panelist-product type-week	Filled with 0
Convergent search:	An observation is a checkout	Not included
No. product pages visited in the transacted category	An observation is a checkout	Not included
No. domains visited in the transacted category	An observation is a checkout	Not included
Total time spent in the transacted category	An observation is a checkout	Not included
Domain traffic (Y_{jt})		
No. page clicks	Domain-week	Filled with 0
No. distinct visitors	Domain-week	Filled with 0
Domain traffic by marketing communication channel	Domain-week	Filled with 0
Fraction of page clicks from ad-click, email, and search	Domain-week	Filled with 0
Fraction of new visitors from ad-click, email, and search	Domain-week	Filled with 0

A.2 Consumer Browsing

Table A.2presents summary statistics (mean and standard deviation) for measures of panelist browsing activities, broken down by country. The average panelist visits between 31.09 to 55.37 domains in a week, spends 533.22 to 778.83 minutes online. On average, a page duration of stay is about 30 seconds.

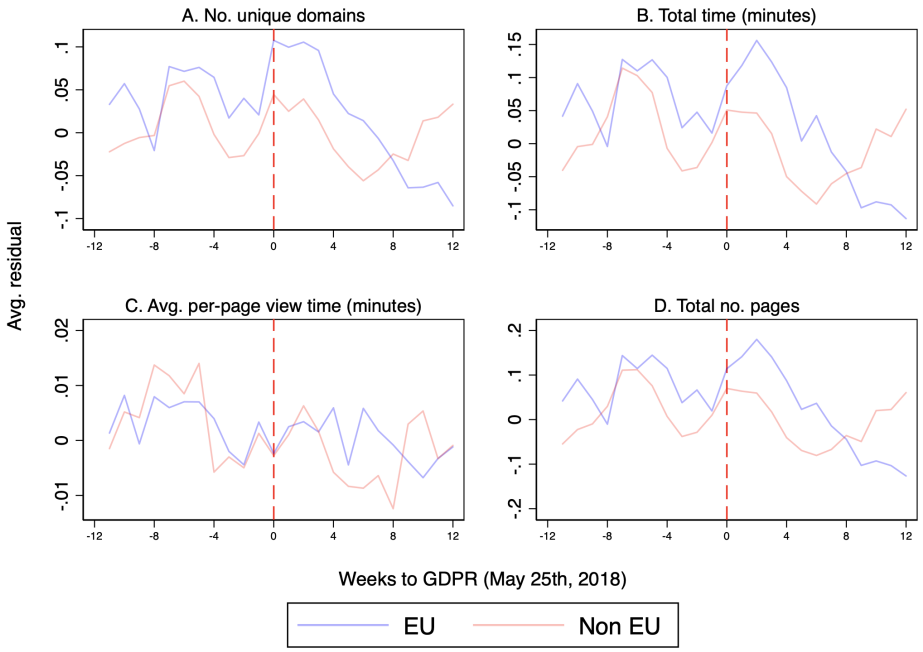
Table A.2: Summary Statistics of Panelist Weekly Browsing Activities

	Mean (std.dev) of			
	US	UK	Spain	Brazil
No. domains	40.79 (51.53)	55.37 (57.71)	41.13 (59.89)	31.09 (38.76)
time (minutes)	657.68 (915.77)	778.83 (905.13)	533.22 (756.17)	562.97 (778.07)
per-page time (minutes)	0.52 (0.66)	0.51 (0.60)	0.46 (0.58)	0.50 (0.61)
No. pages	1048.42 (1576.63)	1447.54 (2166.45)	1042.69 (1941.05)	914.41 (1414.33)
No. obs	51,980	52,992	63,020	70,748

Note: This table summarizes the mean and the standard deviations of panelist weekly browsing activities, broken down by country. Page views shorter than 2 seconds or longer than 12 hours were dropped.

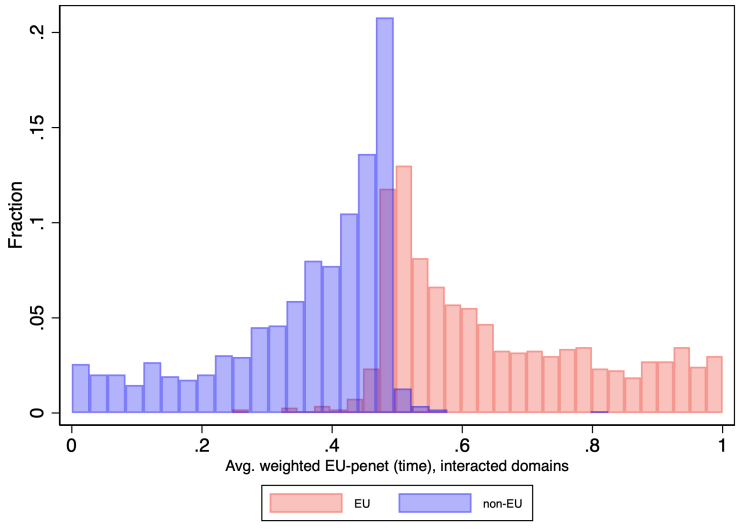
In Figure A.2, we display the average panelist browsing activities, broken down by region over time, for 12 weeks before and after GDPR. The y-axis is the average residual activity: it is obtained by regressing logged outcomes on panelist fixed effects, then by taking the average over all panelists within the same panel. Average domain EU-penetration for EU and non-EU panelists are given in Figure A.3.

Figure A.2: Activities before and after GDPR, Raw data, EU and non-EU panels



Notes: The lines indicate average residual panelist activities, broken down by EU and non-EU. The residuals are obtained by regressing logged outcomes (number of distinct domains visited in A, total time online in B, per-page view time in C, and total pages clicked in D) on panelist fixed effects. Week 0 marks the week of May 25th, 2018, which is the week of GDPR deadline.

Figure A.3: Distribution of panelist average (weighted) EU-penetration by region



Notes: An observation is a panelist. A panelist's (weighted) average EU-penetration is the (weighted) average EU-penetration of all domains he or she interacted with, weighted by the activity (total page clicks, or total view time) at each of the domains.

A.3 Domain Characteristics

To assess the heterogeneous effects of GDPR on large and small firms, we construct measures of domain size using pre-GDPR traffic. We count number of distinct visitors visited a site in the first quarter of 2018 as a proxy for domain size.

GDPR specifies protection for EU citizens only and domains are subject to the requirements of GDPR as long as they process EU citizens' personal information. Therefore, the compliance cost a domain faces depends on how exposed it is to EU customers. To assess to what extent a domain is exposed to EU traffic, we construct a EU-penet$_j$, defined as:

$$\text{EU-penet}_j = \frac{N_j^{\text{EU}}}{N_j}, \tag{4}$$

where N_j^{EU}, and N_j are the number of unique EU visitors and total number of visitors to site j, in the first quarter of 2018.

Table A.3: Summary Statistics of Domain Exposure to EU Traffic and Sizes

Variable	Mean (standard deviation)
EU-penetration	0.54
	(0.48)
No. distinct visitors, pre-GDPR	3.31
	(17.57)
No. obs (domains with more than 1 distinct visitors)	314,353
Employee size (total*)	6,879.95
	(60,159.08)
Max of employee size	2,502.83
	(17,528.79)
Avg. employee size	1,331.65
	(9034.77)
No. obs (domains with employee size)	27,261

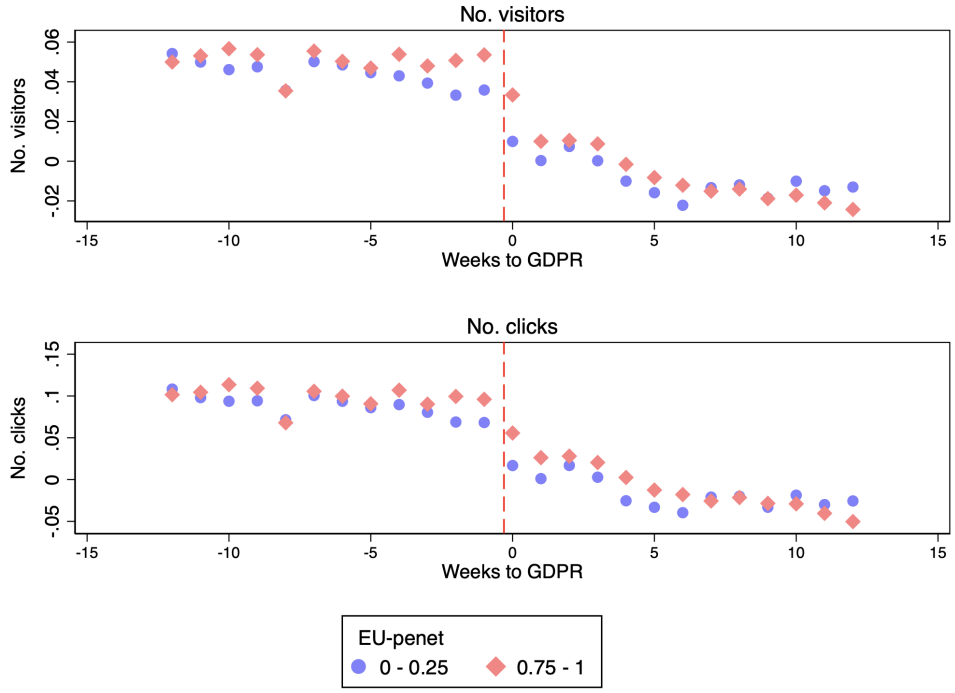
Note: The sample includes domains with at least 1 visitor in the first quarter of 2018. EU-penetration is the fraction of distinct visitors from EU panels among all visitors. There are 314,353 such domains. Among these domains, 27,261 domains have valid employee size information from Bureau van Dijk and/or Crunchbase. For domains with multiple records across different sources, we report the total, max, and average employee sizes.

We also plot the weekly domain traffic, number of unique visitors to the site and total page clicks on a site, for 12 weeks before and after GDPR, broken down by domain's EU exposure (see Table A.3 for its summary statistics). We focus on the top and bottom 25% of the domains as domains with 50-50 EU traffic coincides with the largest sites (such as amazon.com, ebay.com, and so on). While domains in both bins experience a drop in the traffic, the magnitude of such drop is smaller for more EU-exposed domains, consistent with regression results in Table 4.

A.4 Examples of consumer search

Table A.4 provides an example for the consumer search for products and the keywords that they plug in within the same product category, for illustrative purposes. For search terms on the same topic (electronics), UK

Figure A.4: Domain Traffic, by pre-GDPR EU-exposure (top and bottom 25%)



Notes: The y-axis is the residual weekly domain traffic (A. number of distinct visitors, B. number of clicks) from regressing logged outcome on domain fixed effects, and averaged across domains for lower (below 0.25) and higher (above 0.75) EU-penetration.

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.

A.5 GDPR Compliance Dates

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. 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. We obtained update times indicating GDPR compliance for 14,551 websites. Figure A.5 plots the histogram of companies' adopted GDPR policy date.

A.6 Search for GDPR-related Information

We present an illustration of the validity of identified search terms. To see whether the panelists in our sample are aware of the policy enforcement, we count number of search terms containing 'GDPR" (case-insensitive)

Table A.4: Examples of consumer search, US and UK (English-speaking panels), 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			

Note: This table illustrates consumer search by listing the consecutive search keywords submitted with the corresponding timestamps by four panelists from the English-speaking countries (UK and US) on the topic "electronics" before and after the official GDPR date.

in each week for English-speaking panels. Figure B.1 plots the number of times users plugged in the keyword GDPR in their browser over time. There is a spike coinciding with the official GDPR date, May 25, 2018. The peak for UK panel is higher than that of the US panel's.

B Supplemental Information on Methods

In this section, we provide details on the estimators (GSC, MC) used in the analysis. We discuss more details on the hyperparameter selection processes for GSC (number of latent factors for panelist-specific time trends), and how the quality of fit is assessed for both methods.

B.1 Model Selection for Generalized Synthetic Control

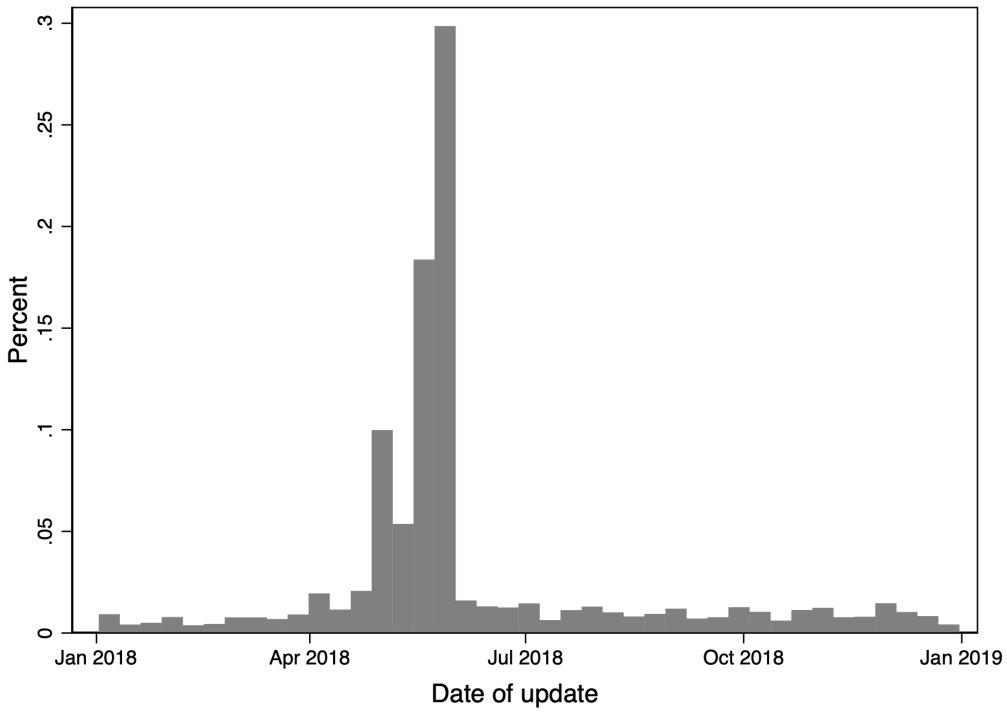
The GSC estimator incorporates individual-specific time trends by estimating individual's factor loadings on r latent factors, common to all panelists. The number of factors, r, is a hyperparameter that needs to be chosen. The heterogeneity across individuals is characterized by individual-specific factor loading for each of the r factors. To choose the optimal hyperparameter, we follow the process suggested by Xu (2017), and choose r that yields the lowest pre-policy fit between the treatment group and its synthetic control / counterfactual.

Specifically, for a given r and corresponding latent time-varying factors $\hat{\mathbf{f}}_t$ estimated from the control group, the following predicted value is computed for all panelists, averaged across periods before the policy deadline:

$$SSE^{\text{Pre}} = \frac{1}{T0}\{\sum_{t=1}^{T0}[\sum_{i\in N0}(\hat{\lambda}_i \cdot \hat{\mathbf{f}}_t) - \sum_{i\in N1}(\hat{\lambda}_i \cdot \hat{\mathbf{f}}_t)]^2\} \tag{5}$$

where $\hat{\lambda}_i$ for the treated panelists are estimated using the same set of $\hat{\mathbf{f}}_t$ as the control group. The expression can be interpreted as the aggregate fit of the model before GDPR, as it is the squared differences between EU outcomes and EU's counterfactual outcomes. Choosing the r that minimizes the SSE therefore results in better parallel pre-trend.

Figure A.5: Distribution of Compliance (privacy policy update) Times



Notes: An observation is a domain. The sample includes domains with identified privacy policy change dates.

B.1.1 Quality of pre-GDPR match

We follow the process given by Liu et al. (2021) to assess the pre-policy fit of the counterfactual estimates. Specifically, the assessment conduct equivalence test on pre-policy treatment effects, and lower p-values of tests indicate better fit. In B.2, we present the results of the tests. We have p-values at 0.00, indicates quality fit of the estimators.

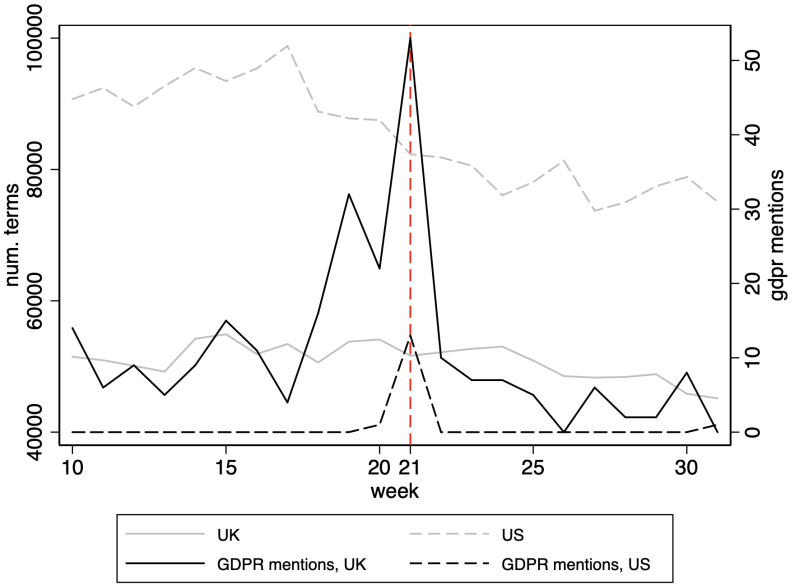
B.2 Panel Differences

In this section, we provide more details about the panel-difference estimator. As we have observations for each panelist in 2019 we also examine a panel difference (PD) estimator, which compares a panelist's activity in 2018 to his or her 2019 activity, week by week. In other words, we follow the specification given in Equation 6:

$$\ln(Y_{ikt}) = \alpha_1 + \alpha_2 \text{GDPR}_t \times \text{Year } 2018_t + \alpha_3 \text{Year } 2018_t + \text{WOY}_t + \epsilon_{it}, \tag{6}$$

where GDPR_t is an indicator, equals to 1 if week t is after the week of May 25th (or the 21st week of a year). Year 2018_t indicates if week t is in year 2018 (as opposed to 2019). We control for week of year fixed effects and panelist fixed effects. Essentially, PD estimator calculates, for the same panelist, the percentage change in her 2018's activity from her 2019's activity for a given calendar week right around the policy enforcement date. The coefficient to the interaction term measures the average treatment effect on the treated (ATT), as we focus on within-panelist comparison for EU panelists. The PD estimator is identified with the assumption

Figure B.1: Number of search queries containing "GDPR", by country



Notes: The lines indicate number of search terms that contain "GDPR" (case-insensitive), submitted by US and UK panelists. Week 21 is the week of May 25th. The total number of search terms submitted by US and UK panelists are plotted in the back.

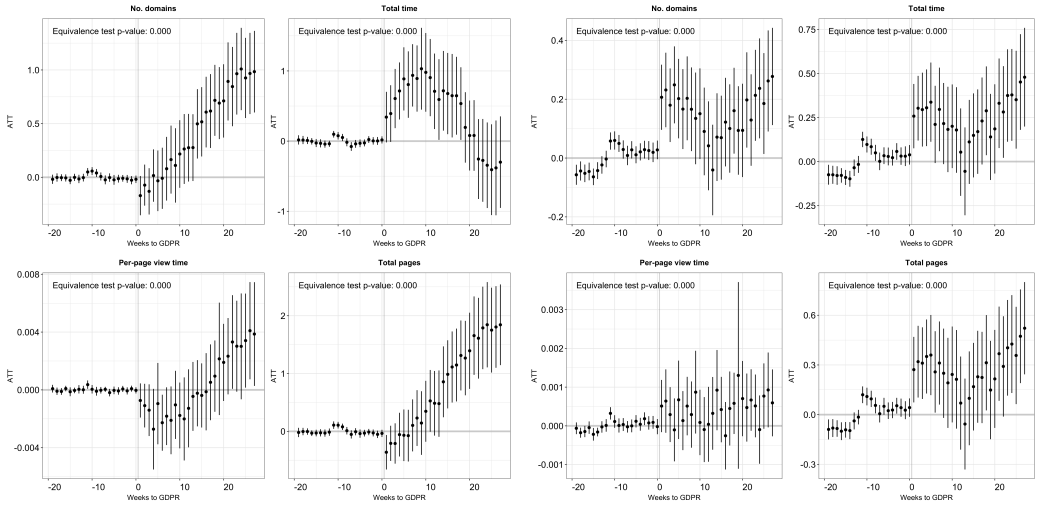
that around the GDPR date of 2019, panelists should not exhibit the same changes in their online behavior as they did in 2018. By examining the differences between 2018's and 2019's activities, we are also eliminating individual-specific time trends that might confound GDPR's effects.

B.3 Search Term Pre-processing

We pre-process the search terms following several steps. First, the search keywords are identified by looking for URL parameters indicative of query use. These URL parameters are "query", "q", "keyword", the same query is recorded every time the panelist browses an additional page of results. This creates several duplicate search terms, which indicate a longer time examining results, but does not indicate that the panelist is starting a new search with the same keyword. For our topic modelling, we dropped, for each panelist, all redundant queries under the same domain in the same hour. This allows us to capture more topics, where smaller topics (in number of search terms) could have been grouped into one topic if the largest ones are dominated by repeating terms. However, as repeated search terms indicate panelists browsing more result pages, such observation indicates increase in search effort, and we keep the duplicate terms in the robustness check where we define search episodes by looking at jumps in semantic similarity.

Next, we dropped the identified terms 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, where a word is

Figure B.2: Synthetic Control pre-GDPR Fit, GSC (left) and MC



Notes: The plots presents estimates and 95% confidence intervals for ATTs over weeks from GSC and MC estimates. We report equivalence test p-value, where lower p-values indicate better pre-GDPR fit. All estimates include individual specific time trends. For both estimators, we use 10-fold cross-validation to choose the hyperparameters, based on the pre-policy fit.

a sequence of letters without any delimiters in-between. 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.

B.4 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:

$$\mathcal{LL} = \sum_{j}^{W} \{ \sum_{i:\ i,j} log(P(D_{ij} = 1|v_i, v_j)) + \sum_{i\prime \sim U(w)}^{2} log(1 - P(D_{i\prime j} = 1|v_{i\prime}, v_j))) \ \} \tag{7}$$

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.

We identify latent topics using a skip-gram model jointly with k-means clustering. Skip-gram model predicts the set of words that are most likely to appear next to a particular word w in a text (Mikolov et al., 2013; Levy and Goldberg, 2014). More specifically, the probability that the words w_i and w_j can appear within c words of each other in a text can be written as:

$$P(D_{ij}^c = 1|w_i, w_j) = \frac{\exp(v_i^\top v_j)}{\sum_{t=1}^{W} \exp(v_t^\top v_j)} \tag{8}$$

where D_{ij}^c takes the value 1 when words i and j co-appear in the specified window c. We set c to two since the average search term in the data contains 5.2 words. In the above equation, the vector v_i is the word embedding – a numeric vector – that indicates word i's location in the latent semantic space. Word embeddings characterize similarities between words (Timoshenko and Hauser, 2019; Arora et al., 2017): a pair of words share common contexts in the data are similar in meaning and will have word embeddings that are close to each other, where closeness is measured by cosine similarity. For example, the three most similar words to the word "car" are "vehicle," "truck," and "lease" in the US keyword data and "garage," "property," and "vehicle" in UK keyword data. We obtain the vector representation of a search term by taking the average of words' embeddings contained in that search term. More details are provided in Appendix B.4.

Finally, we perform k-means clustering on the search term vectors (Hashimoto et al., 2016; Cha et al., 2017) to assign each term to a cluster, which then becomes a latent topic. Let v_i be a search term's embedding, then its topic, k_{v_i}, is identified using k-means clustering:

$$k_{v_i} = \underset{k\prime}{argmin} ||c^{k\prime} - v_i|| \tag{9}$$

where $c^{k\prime}$ is the centroid of cluster $k\prime$; $||c^{k\prime} - v_i||$ is the Euclidean distance between term v_i and the centroid. This is done in a way to minimize the sum of squared distances of objects to the cluster centroid, where distances are measured as Euclidean distance. Using the elbow point in the total squared distance as a rule-of-thumb to decide how many clusters to keep results in 50 clusters for the UK panel and 65 clusters for the US panel. Examples of top 10 topics from each panel and related search queries are provided in

For illustration, we show examples of the sequential search terms submitted, and how we identify the products viewed and checkouts that follow, if any.

Figure B.3 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 Figure B.3, 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.

Figure B.3: 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**	

A checkout

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**

Notes: This figure is a snapshot of a panelist's browsing records under the "clothing and accessories category". We identify the product category as the phrase, "jewelry-pendants-necklaces" falls under "clothing and accessories" according to Google Product Taxonomy.

C Supplemental Analysis

C.1 Alternate Measure of Search Efforts

To complement the results in Table 1, we exploit the "streams of search queries" observed in our data, and identify the start of a new search episode based on inter-query similarity: once the next query's similarity is lower than the 25th percentile of any inter-query similarity submitted by that panelist, we mark that as the start of the next episode. This way, we split the search streams into episodes at "jumps" in semantic similarity, without imposing additional assumptions about the number of latent topics. We focus on a series of measures of search efforts at "episode" levels: number of search queries used in, time span (i.e., the time gap between first and the last term from the same episode) of an episode, page view time for each of the result pages, and total view time on the search engine result pages. We also examine whether the number of episodes increase after GDPR.

The main difference of this measure, compared to the one we use in the paper is that a panelist may search for the same topic in multiple episodes, and between two of such episodes, there may be terms submitted under different topics. For example, a consumer may search for car insurance for multiple days, while searching for restaurants in between. This alternate measure of search effort would bias the actual effort exerted on "car insurance" downwards. The other disadvantage is inability to control for topic fixed effects. We therefore rely

on the search effort as our preferred measure of consumer search length and view this episode-level effort as complementary.

Table C.1 reports the estimates of the interaction term, GDRP × EU, for the five outcomes discuss above. The results imply a 0.3% increase in the number of queries used per episode, 3.9% increase in the time span of an episode, 0.7% increase in the number of episodes per week. The time spent inspecting the result pages also increased for EU users: for each result page, the page view time increases by 0.3%, and the total time spent on all result pages increased by increase by 2.1%. These results are consistent with EU panelists putting more efforts into search after GDPR.

Table C.1: Change in Search Efforts (per Search Episode)

	(1) No. queries	(2) Time span	(3) Result page view time, total	(4) No. episodes, week	(5) Result page view time, each
GDPR × EU	0.007^{***}	0.052^{***}	0.009^{***}	0.015	0.003^{***}
	(0.002)	(0.008)	(0.002)	(0.012)	(0.001)
No. obs	1,171,852	1,171,852	1,171,852	69,608	2,135,983
Panelist FE	Yes	Yes	Yes	Yes	Yes
Day of week FE	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes
Mean of DV	0.938	0.723	0.354	2.246	0.228
Std.dev of DV	0.392	1.741	0.385	1.066	0.269

Note: In columns (1), (2), and (3), each observation is a search episode. In column (4), each observation is a panelist-week. In column (5), each observation is a search result page view. The start of a search episode is marked by a lowered cosine similarity between two consecutive terms: if a query's similarity to the previous one is below the 25th percentile of all inter-query similarity of that panelist, this query is considered to be the start of the next episode. $\log(x+1)$ transformation is applied to all outcomes. Heteroscedasticity-robust standard errors clustered at the domain level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$