Who Wins, Who Loses under Privacy Regulation? Evidence from the GDPR and Online Tracking

Gabriel Solis*

Department of Economics, University of Southern California

Advised by Jonathan Libgober

Spring 2025

Abstract

This paper asks whether the European Union's General Data Protection Regulation (GDPR) changed online tracking and the concentration of the tracking market. I use public WhoTracks.Me data from May 2017 through February 2025. A major challenge is that the way these data were collected changed around the same time GDPR took effect: Ghostery users began contributing in early 2018, the contributor base became more international, and new regional files were introduced. Because of these changes, I treat the results cautiously and distinguish changes in the recorded data from effects that can be attributed to GDPR. For the website analysis, I use the global WhoTracks.Me series, compare websites with EU and non-EU country-code domains, exclude the March–May 2018 transition period, and include website and month fixed effects. I find no statistically significant decline in tracking after GDPR. The estimated post-period change is +1.05 tracking services per website and −0.01 in the share of page loads with detected identifier transmission; neither estimate is statistically distinguishable from zero. The tracking outcomes also show different trends before GDPR, which makes a causal interpretation difficult. I also examine concentration after grouping tracking services by their parent companies. Average monthly HHI rises from 331 before the transition to 508 afterward, but the increase begins before GDPR became enforceable. Nearly all of the exact HHI increase comes from shifts among companies that appear in both periods, especially an increase in Google's relative observed reach, rather than the disappearance of smaller firms. The results do not provide clear evidence that GDPR caused a privacy–competition trade-off. Instead, they show how conclusions from online tracking data can depend heavily on how the data are collected, how treatment is defined, and how tracking firms are measured.

*Email: gabe.solis504@gmail.com. I am deeply grateful to Professor Jonathan Libgober, whose guidance and first-year Ph.D. microtheory course opened new directions for this research. I also thank the USC Economics Department, whose faculty's teaching, support, and opportunities empowered me to pursue my own questions. Special thanks to Professor Ergin Bayrak for his unwavering mentorship and dedication to my growth as a researcher. Any remaining errors are my own.

Contents

1 Introduction **1**

2 Background and Related Evidence **3**

2.1 GDPR . . . 3

2.2 Enforcement as institutional context . . . 3

2.3 Online tracking and ownership . . . 4

2.4 Related empirical evidence . . . 5

3 Data and Measurement **7**

3.1 WhoTracks.Me data layers . . . 7

3.2 The data-generating process . . . 8

3.3 Website outcomes . . . 9

3.4 Domain classification and preferred sample . . . 10

3.5 Parent-company concentration . . . 10

4 Empirical Framework **11**

4.1 Website fixed-effects comparison . . . 11

4.2 Inference and sensitivity analysis . . . 11

4.3 Descriptive concentration analysis . . . 12

5 Results **12**

5.1 Raw website trends . . . 12

5.2 Preferred website estimates . . . 13

5.3 Measurement and composition diagnostics . . . 15

5.4 Sensitivity across specifications . . . 16

5.5 Parent-company concentration . . . 16

5.6 Why ownership aggregation matters . . . 18

5.7 Decomposing the HHI increase . . . 19

6 Discussion **19**

7 Conclusion **21**

A Supplementary Data, Definitions, and Sample Construction **22**

B Supplementary Results and Diagnostics **24**

1 Introduction

Regulation shapes the trade-off between consumers and firms in an exchange of information, so privacy is as much an economic question as a legal one (Acquisti et al., 2016). The General Data Protection Regulation (GDPR) introduced new legal rules for collecting and processing personal information in Europe, and it also generated an interesting empirical puzzle. Privacy regulation can reduce unwanted data collection, but compliance is often harder for small firms than for large rivals. If both are true, GDPR could reduce tracking overall while still shifting it toward the largest platforms. This possibility comes up in theoretical work on privacy regulation and market structure (Campbell et al., 2015) and in recent empirical work on GDPR (Aridor et al., 2023; Johnson et al., 2023). It is hard to test empirically, though, because the data used to measure online tracking are themselves produced by browser extensions, classification rules, and changing samples of users.

I frame my empirical analysis around three research questions:

Research Question 1 (Privacy and measurement): *Did recorded third-party tracking change differentially after GDPR among websites using EU country-code top-level domains, relative to supported non-EU domains?*

Research Question 2 (Market structure): *How did concentration in observed tracking change around 2018 once tracking services are aggregated to their parent companies?*

Research Question 3 (Identification and mechanism): *How sensitive are these results to the treatment definition, sample composition, pre-period trends, and the level of ownership aggregation, and which companies mechanically account for any increase in concentration?*

The distinction between "recorded" and "actual" tracking matters here. WhoTracks.Me measures third-party activity as reported by its participants, so it does not see every page load, every website, or every tracking technique on the internet. The monthly files are valuable because they are large and detailed, but a lot of records does not by itself explain how they were generated. The three research questions call for different kinds of evidence. The website comparison is a difference-in-differences design in a restricted sample. The concentration analysis is descriptive, since there is no stable geographic counterfactual for concentration. The diagnostic question just asks whether either result holds up against the main threats created by how the data were collected.

The key identification problem here is timing. Prior to 2018, WhoTracks.Me traffic came exclusively from users of the Cliqz browser and extension. Ghostery users began contributing in March–April 2018, which increased the size and geographic diversity of the contributor base. WhoTracks.Me itself reports that the average number of detected trackers dropped in April 2018, because Ghostery users blocked more trackers than Cliqz users had. The regional EU and U.S. monthly files also start appearing in March–April 2018 (WhoTracks.Me, n.d.). GDPR came into effect on May 25, 2018. So a simple before–after comparison ends up mixing a legal change together with a change in who was contributing the data.

My first contribution is to rebuild the website analysis around this timing problem. I use the global file, since it is the only one available both before and after GDPR. Websites are classified by their country-code top-level domain (ccTLD), the suffix at the end of a domain name that is assigned to a country or territory, such as `.de` for Germany or `.fr` for France. I compare EU-28 country-code top-level domain websites with supported non-EU country-code top-level domain websites, omit March–May 2018, and include only websites that appear in the data both before and after the transition. The preferred specification includes website and calendar-month fixed effects, with standard errors clustered by country-code top-level domain group and month. I also estimate an equal-weight country-month version of the model using a wild-cluster bootstrap, as a check on the clustering assumption. These choices make the comparison clearer, but they do not turn it into a nationally representative EU sample. The supported sample contains only ten country-code top-level domain groups, and 361 of the 399 treated websites are `.de` sites. So I read these estimates as a restricted differential comparison, not as the average effect of GDPR on all EU websites.

The preferred estimates do not support a significant drop in tracking. The point estimate for the number of tracking services per website is +1.05 ($SE = 0.95$). For the share of page loads with any detected identifier transmission, the estimate is −0.010, with a standard error of 0.019, and cookie-based tracking looks similarly close to zero. The 95 percent confidence interval for tracker count runs from −0.80 to +2.90 services, relative to a pre-period EU-domain mean of 12.11, and it excludes a reduction as large as the one found in the original pooled hybrid-classification specification. The direct tracking-prevalence outcomes also show positive slopes before the transition, which weakens a parallel-trends interpretation of the results.

My second contribution is to rebuild market concentration at the parent-company level. WhoTracks.Me identifies individual tracking services separately from the companies that own them, so a service-level calculation would treat Google Analytics, DoubleClick, and other Google services as separate entities. That's useful for measuring the diversity of tracking technology, but it's the wrong level for measuring ownership concentration. I use the official monthly `companies.csv` files and define each company's share as its reach divided by the total reach of all companies in the same month. Because several companies can appear on the same page load, these normalized shares measure relative tracking presence, not revenue or conventional market share.

Concentration rises sharply in early 2018. The average monthly Herfindahl-Hirschman Index (HHI) rises from 331 in May 2017–February 2018 to 508 in June 2018–May 2019, and the top-four share rises from 27.5 to 33.2 percent. The timing does not fit a simple GDPR story, though: HHI rises from 319 in December 2017 to 378 in January 2018 and 462 in February, before GDPR becomes enforceable and before the contributor transition is complete. The data also rule out a simple exit-based story. The number of observed companies goes up over this period, not down, and companies present in both windows account for 99.2 percent of the exact HHI increase. Google's normalized reach share rises from 14.78 to 19.94 percent and accounts for nearly all of the mechanical increase in HHI. This is a result about Google's relative observed reach, not proof that GDPR caused Google to gain market power.

My conclusion is narrower than the question in the title. The data show which companies gained relative presence in the published tracking records. That is not the same as identifying causal winners and losers from

GDPR. This changes what I think counts as good empirical work here. A precise coefficient estimated from millions of observations is not, by itself, convincing. What matters more is whether the policy variation can be separated from changes in measurement, and whether the outcome being measured actually lines up with the economic quantity under discussion. Uncertainty here should also reflect the level at which treatment varies, not just the number of underlying observations.

The rest of the paper is structured as follows. Section 2 gives background on GDPR, the online tracking ecosystem, and related literature. Section 3 explains the data and the measurement choices made here. Section 4 presents the empirical framework. Section 5 reports the website and concentration results. Section 6 discusses what this evidence can and cannot establish. Section 7 concludes.

2 Background and Related Evidence

2.1 GDPR

GDPR became enforceable on May 25, 2018. It applies to the processing of personal data by organizations established in the European Union and, under certain conditions, to organizations outside the EU that offer goods or services to individuals in the EU or monitor their behavior. It requires a lawful basis for processing, and when the basis is consent, that consent must be freely given, specific, informed, and unambiguous. GDPR also includes principles of purpose limitation, data minimization, and accountability, along with rights of access, erasure, and portability. Administrative fines are capped at €20 million or 4 percent of annual worldwide turnover, whichever is greater (European Parliament and Council of the European Union, 2016).

Two things about GDPR matter for how I set up the empirical comparison. One is that legal exposure isn't defined by a website's country-code top-level domain. A `.com` website can still be subject to GDPR, and a `.de` website may serve users who are entirely outside the EU. The other is that GDPR can affect the behavior of organizations that aren't themselves based in the EU, through shared technical infrastructure and voluntary compliance. This broader impact is sometimes called the Brussels Effect (Bradford, 2020), and the related empirical literature reports spillovers in website behavior beyond the most directly regulated group (Peukert et al., 2022). So a non-EU comparison group can end up partially exposed to GDPR as well.

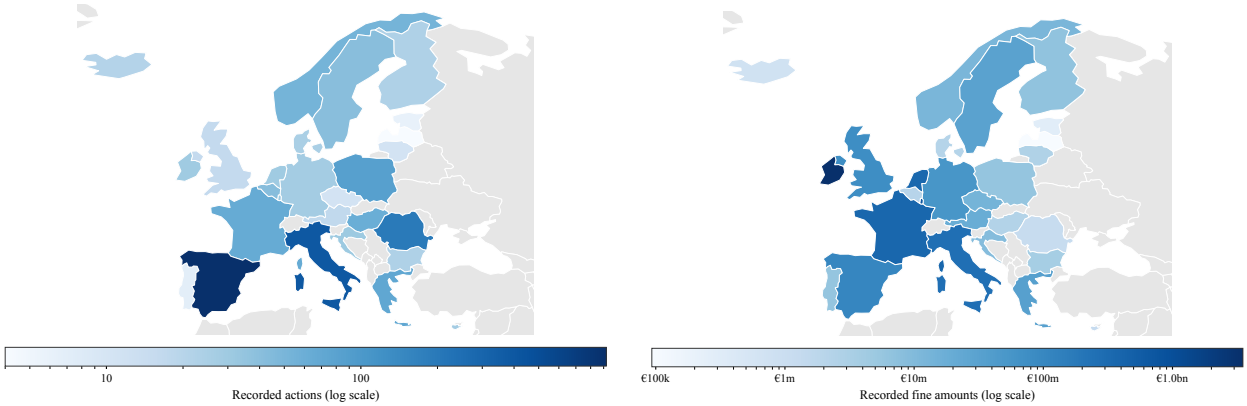
It follows that GDPR exposure can't be assigned as a clean binary treatment. I treat the domain classification as a proxy for likely geographic orientation, not as a legal determination. The empirical question is whether recorded outcomes differ for EU-related domains, given these limits on what the identification strategy can actually pin down.

2.2 Enforcement as institutional context

GDPR provides a unified legal framework, but individual investigations and sanctions are carried out by national data protection authorities. Publicly recorded enforcement varies a lot across countries as a result. Figure 1 shows two distinct sides of that variation. Spain and Italy account for a large share of recorded

enforcement actions, but the euro value of penalties is dominated by a small number of countries and an even smaller number of exceptional cases. The maps provide useful institutional context. They are not a measure of exogenous enforcement intensity, though: fine counts depend on reporting practices, the mix of cases each authority pursues, available resources, and where large firms happen to be located, and fine amounts are especially sensitive to a handful of landmark decisions. I don't use fines to define treatment. I also don't use them to estimate heterogeneous effects or to explain the website results.

Figure 1: Publicly Recorded GDPR Enforcement and Fine Amounts by Country



(a) Number of publicly recorded actions

(b) Cumulative recorded fine amounts

Notes: The maps aggregate the country-month enforcement extract constructed from 2,560 public records in the GDPR Enforcement Tracker, including decisions recorded through February 20, 2025 (GDPR Enforcement Tracker, 2025). Panel (a) sums recorded enforcement actions by country; Panel (b) sums reported euro amounts. Both color scales are logarithmic because the distribution of values across countries is highly skewed. Gray areas have no matched positive value. Counts and amounts are descriptive and are not used as a treatment or enforcement-intensity measure. Country boundaries are from Natural Earth (Natural Earth, 2025).

2.3 Online tracking and ownership

A website is a first party: it is the domain that the user chooses to visit. Other domains embedded on the page are third parties. Third parties may provide analytics, advertising, content delivery, social-media widgets, fraud prevention, and other services, and some of them send identifiers that link activity across pages or across websites. WhoTracks.Me groups known third-party domains into tracking services, and it separately groups those services under parent companies, using TrackerDB (Karaj et al., 2018; WhoTracks.Me, n.d.; Yu et al., 2016).

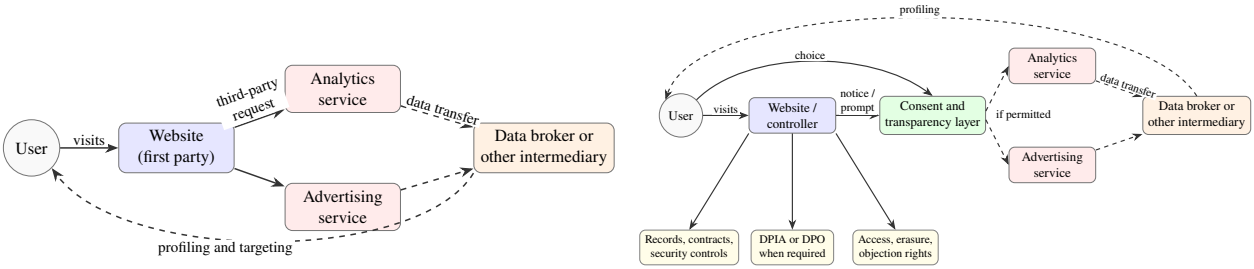
A tracking service is a technology or product identifier. A parent company is the organization that owns one or more of these services. Keeping the two separate lets me distinguish the concentration of technologies from the concentration of the organizations that actually control them, and it's the second object that's closer to the market-structure question in my title.

Data create scale advantages for firms, since more observations of user behavior improve targeting, measurement, or product design, so the value of data can be dynamic and nonrival (Hagiu & Wright, 2023; Jones & Tonetti, 2020; Prüfer & Schottmüller, 2021). There are a few different ways privacy rules could affect these

scale advantages. Rules might restrict data collection equally across all firms, impose compliance costs that fall harder on smaller firms, or strengthen firms that already have direct customer relationships and large first-party data stocks (Campbell et al., 2015). Which of these dominates is an empirical question.

Figure 2 gives a stylized picture of the economic relationships behind these outcomes. Before any consent choice is made, third-party analytics and advertising services can load on the page a user visits, send identifiers or event data, and feed into profiles built across websites. Under a consent-based GDPR workflow, nonessential flows depend on the user's choice and operate within a broader governance system. The figure separates the user-facing consent interface from accountability obligations and user rights, since GDPR does not make consent the only lawful basis for processing. Requirements such as a Data Protection Impact Assessment or a Data Protection Officer apply only under specified conditions.

Figure 2: How Online Tracking Data Flows, Before and After GDPR



(a) Data flow with no consent step in between

(b) Consent-based tracking under GDPR

Notes: This figure is a simplified illustration. It does not capture every detail of EU privacy law or of any single website's actual setup. Panel (a) shows third-party services loading automatically, with no consent step involved. Panel (b) shows a consent-based workflow instead, where data only reaches third parties if the user agrees, and where the process also involves transparency requirements, purpose limitation, data minimization, security safeguards, recordkeeping, and user rights such as access and erasure. GDPR does not require consent for every use of data; organizations can rely on other lawful bases too. A Data Protection Impact Assessment (DPIA) or a Data Protection Officer (DPO) is required only in specific circumstances, and the right to erasure is a separate user right, not part of the consent step itself.

The figure also shows the limits of what the WhoTracks.Me project actually records. It includes only third-party presence and selected identifier signals. It does not include the lawful basis for processing, the text or validity of consent, the purpose of processing, or whether data-subject requests are honored. A decline in observed tracking could reflect real compliance with GDPR, but it could just as easily come from more blocking or from a change in who is contributing data. A stable tracker count doesn't prove GDPR had no effect, either, since firms could have changed what data they collect or how they use it in ways that don't show up as third-party tracking at all.

2.4 Related empirical evidence

Economic analysis of privacy rules starts from a trade-off: personal data can improve matching, measurement, and product design, but consumers bear costs from unwanted disclosure and from uses of their data they cannot observe (Acquisti et al., 2016). In advertising markets, privacy rules can reduce the effectiveness of behavioral targeting (Goldfarb & Tucker, 2011). In market-structure models, conditioning data collection on consent can favor firms that already have broad customer relationships or large first-party data stocks (Campbell et al., 2015). These mechanisms motivate my empirical question, but they don't predict its sign

on their own.

Empirical work on GDPR looks at fairly different units and margins depending on the study. Aridor et al. (2023) use consumer-level data from an online travel intermediary and find fewer observable consumers after the consent requirement, but greater persistence and value among the consumers who remain observable. Godinho de Matos and Adjerid (2022) study consent and targeting inside a telecommunications firm. Settings like these let researchers observe choices or firm behavior more directly than a public website panel does, but they also cover narrower markets.

A second group of papers studies websites and technology providers directly, which is closer to what I do here. Peukert et al. (2022) follow more than 110,000 websites and document fewer interactions with web-technology providers, spillovers to websites outside the directly regulated group, and rising concentration, since Google loses less reach than smaller providers do. Johnson et al. (2023) use vendor inventories for more than 27,000 leading websites and find a sharp short-run decline in vendor use for EU users, along with higher relative vendor concentration. Their vendor-level data are close to my tracking-service layer. My ownership analysis goes a step further and consolidates services under parent companies.

One working paper that also uses WhoTracks.Me is especially relevant. Miller et al. (2024) use a balanced panel of 294 publishers and report about four fewer trackers per publisher after GDPR, a 14.79 percent relative decline. My analysis is not a direct replication of that estimate. I use only the global series and define exposure using supported country-code top-level domain groups, not the full panel of publishers. I also drop the contributor transition period from the sample and cluster standard errors at the domain-group level. And I pay closer attention to changes in the data-generating process itself. I think the contrast is useful, since it shows how much conclusions drawn from the same underlying data can depend on which records are selected and which unit of analysis is used.

Another group of papers focuses on innovation and investment rather than tracking itself. Janßen et al. (2022) find a decline in app availability around GDPR, and Jia et al. (2021) find lower short-run technology venture investment in the EU. These outcomes help inform the larger policy discussion, but they don't validate any particular tracker-count design.

Table 1 compares the closest empirical work. The studies differ in unit of observation, policy exposure, horizon, and measurement system. My contribution is to make those differences central to the comparison. I examine what WhoTracks.Me data can actually say about GDPR, and I distinguish the concentration of tracking technologies from the concentration of the companies that own them.

Table 1: How This Study Relates to Selected Empirical GDPR Research

Study	Data and empirical comparison	Main finding and relation to this paper
Aridor et al. (2023)	Consumer-level data from an online travel intermediary; EU and non-EU markets around GDPR.	The number of intermediary-observed consumers falls by 12.5 percent. Consumers who remain observable are tracked for longer and become more valuable to advertisers. The unit here is an identifiable consumer, unlike the website-level tracking outcomes studied in this paper.
Peukert et al. (2022)	More than 110,000 websites and their third-party HTTP requests, observed for 12 months before and six months after GDPR took effect.	Websites reduce interactions with technology providers, including websites outside the directly regulated group. Google loses less reach than smaller providers do. Their data and treatment construction differ from the browser-contributor panel studied here.
Johnson et al. (2023)	Technology vendors used by more than 27,000 leading websites internationally; short-run comparison around May 25, 2018.	Vendor use falls for EU users, and relative vendor concentration rises, especially among data-intensive vendors. Their vendor inventory is closer to my service layer. My concentration analysis goes a step further and aggregates services up to parent companies.
Miller et al. (2024)	Balanced panel of 294 publishers from WhoTracks.Me, May 2017–December 2019.	This working paper reports about four fewer trackers per publisher, a 14.79 percent relative decline. I use the same underlying project, but I focus on the March–April 2018 contributor transition, a global-series country-code top-level domain comparison, and inference based on ten supported domain groups.
This paper	WhoTracks.Me website, service, site–service, and official parent-company files; May 2017 onward.	The preferred website estimates are small or imprecise, and the direct prevalence measures show differential pre-trends. Parent-company concentration rises descriptively before enforcement, driven mainly by Google's relative reach rather than by small companies disappearing.

Notes: These studies estimate different objects, using different units, exposure definitions, and observation systems, so their numerical estimates are not directly comparable to each other. The table is meant to show how data provenance and the level of aggregation shape what each estimate can actually be interpreted to mean.

3 Data and Measurement

3.1 WhoTracks.Me data layers

WhoTracks.Me publishes monthly aggregates built from users who contribute browsing data through participating software. The project provides five main files in total: three website-related files (`sites.csv`, `sites_trackers.csv`, and `domains.csv`), and two files describing trackers and their owners (`trackers.csv` and `companies.csv`) (WhoTracks.Me, n.d.). I use four of these five files: `sites.csv`, `sites_trackers.csv`, `trackers.csv`, and `companies.csv`.

The website file reports outcomes for popular first-party websites. The website–service file links a site to each classified tracking service observed on that site. The service file aggregates domains attributed to the same tracker, and the company file aggregates services further, up to the parent organization that owns them. Metrics in the website file cover all observed third-party activity, while the service and company files only cover third parties with an entry in TrackerDB. So these layers answer related questions, but not quite the same question.

Table 2 summarizes the cleaned files. The website panel contains 2.73 million website-by-region-by-month records. The website–service file contains 61.9 million relationships, which I process in chunks and aggregate before estimation. The official parent-company files add 80,837 company-by-region-by-month observations, spanning May 2017 through May 2020. This is a longer window than the preferred website regression sample described below, since the concentration analysis draws on the full run of published company files rather than the narrower estimation sample built for the privacy design.

Table 2: The WhoTracks.Me Files Used in This Analysis

Data layer	Records	Unit and role
Website panel	2,730,986	Website-by-region-by-month records used to construct tracking outcomes and the preferred global-series sample.
Tracker-service panel	337,135	Tracking-service-by-region-by-month records used to compare service-level and ownership-level concentration.
Website–service links	61,897,801	Website-by-tracking-service-by-region-by-month records used for aggregation and measurement checks.
Parent-company panel	80,837	Parent-company-by-region-by-month records from the official `companies.csv` files, used to measure ownership concentration.
Preferred website sample	8,941	Website-month observations from 425 websites across 10 supported country-code top-level domain groups, used in the main fixed-effects estimates.

Notes: The underlying WhoTracks.Me files span May 2017 through February 2025. The preferred website analysis, though, uses only May 2017 through May 2019 and omits March–May 2018. The parent-company panel spans May 2017 through May 2020, a separate and longer window used for the concentration analysis rather than the website regressions. Record counts here refer to the cleaned files used in this paper, not the raw downloads. The website–service file is large enough that I aggregate it before running any estimation on it.

3.2 The data-generating process

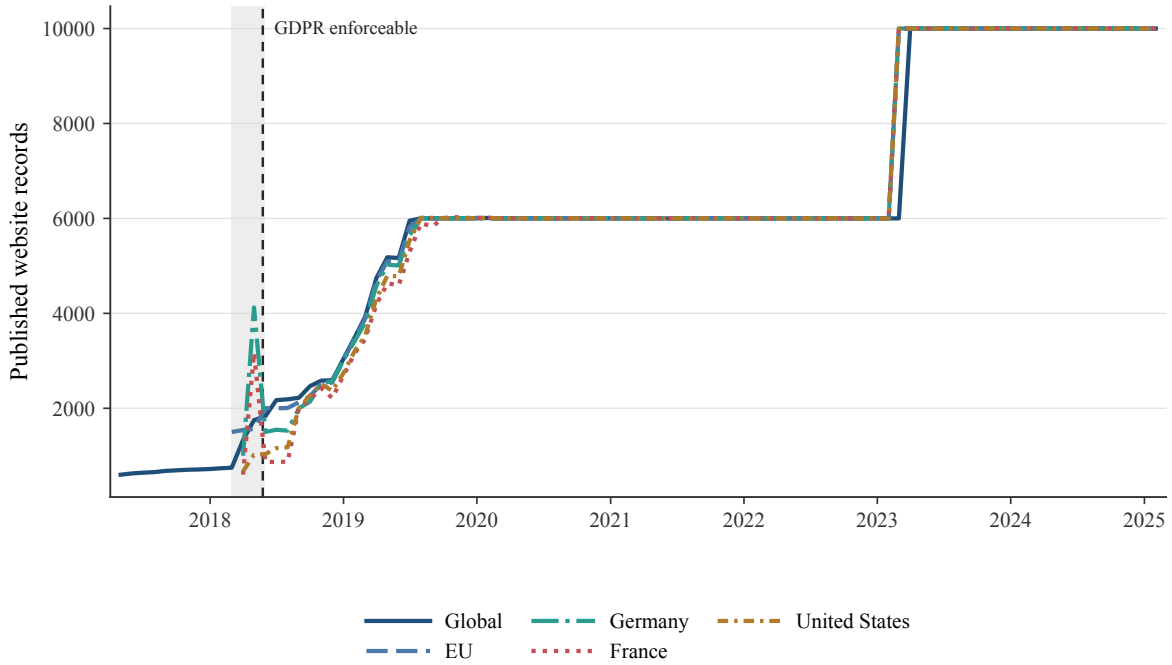
The main limitation here isn't file size. It's a change in where the observations actually come from. Before 2018, all traffic came from users of the Cliqz browser and extension, and WhoTracks.Me reports that roughly 70 percent of its February 2018 observations came from German users alone. Ghostery Firefox users started contributing in March 2018, and users on other browsers followed shortly after. This made the sample larger and more international almost overnight. WhoTracks.Me also reports a slight drop in average tracker counts in April, since Ghostery users blocked more trackers than Cliqz users had. Regional EU, German, French, and U.S. files all begin appearing in March–April 2018 (WhoTracks.Me, n.d.).

The set of published websites changes too. The project describes its site file as a list of popular websites based on contributor traffic, and the number of published global sites rose from roughly 700 in 2017 to more than 1,000 by 2020. WhoTracks.Me also notes that pages without any detected trackers aren't reported at all. So this isn't a fixed census of the web sitting still for me to observe. A change in the numbers can come from actual website behavior, but it can just as easily come from who's contributing data that month, how much blocking is happening, where the reporting threshold sits, or how TrackerDB classifies a given tracker.

Figure 3 makes this coverage change visible. Only the global website series runs continuously from May

2017 through February 2018. An EU series shows up starting in March 2018, and German, French, and U.S. files appear the following month. Records in several of these files then climb quickly toward a shared publication ceiling. None of these regional series, in other words, gives me a long, consistently measured pre-period to work with. If I treated a newly appearing regional label as though it had existed under the same measurement system all along, I'd be confounding geography itself with the simple fact that the file was just created.

Figure 3: How WhoTracks.Me's Geographic Coverage Changed Over Time



Notes: This figure plots the number of published website records in each monthly geographic file. These files are separate views of contributor traffic, not pieces of one combined total, so they should not be summed together. The shaded band marks March–May 2018, the period when Ghostery contributors and the new regional files entered the data and GDPR became enforceable. The dashed vertical line marks May 25, 2018, the date GDPR took effect.

These facts drove how I set up the analysis. I don't use the regional labels for the preferred privacy design, since they barely exist before GDPR takes effect. I omit March–May 2018 entirely, since that's the window where contributor composition and the legal regime are both shifting at once. And I report raw trends and pre-period diagnostics before I try to interpret any post-period coefficient.

3.3 Website outcomes

The main count outcomes are the number of tracking services and parent companies detected per website. The main prevalence outcome, `tracked`, is the proportion of observed page loads on which WhoTracks.Me detects identifier transmission through either a cookie or a query-string parameter. `cookies` is the proportion of page loads on which a cookie is sent by the browser or a `Set-Cookie` response is returned. The original variable `bad_qs` is the proportion of page loads with an identifier-like value detected in the query string. I call this query-string identifier transmission rather than browser fingerprinting, since the documentation

itself doesn't define it as a general fingerprinting measure (WhoTracks.Me, n.d.).

Each prevalence measure lies between zero and one. A value of 0.80 means the feature is detected on 80 percent of the observed page loads represented by that website-month record. These measures depend on which pages and which contributors happen to be in the sample that month. They don't reveal whether the same individual was tracked across time, and they don't reveal whether the underlying processing complied with GDPR. Appendix Table 6 collects formal definitions for these outcomes and for the concentration measures introduced later in this section.

3.4 Domain classification and preferred sample

A country-code top-level domain is the country or territory suffix at the end of a domain name, such as `.de`, `.fr`, or `.ch`. I extract this suffix from each website in the global file and define the treated group as domains associated with the 28 EU member states as of May 2018. The United Kingdom is treated as EU here, since it was still a member at that date. I also include `.eu`, the regional suffix associated with the European Union itself.

The comparison group is non-EU ccTLDs, after excluding European Economic Area countries and suffixes commonly marketed for generic uses, such as `.io`, `.ai`, `.tv`, and `.me`. This exclusion cuts down on obvious classification errors, though it doesn't make ccTLD a perfect measure of geographic location. I keep a top-level-domain group only if it has at least five pre-period months, six post-period months, ten pre-period website-month records, and ten post-period records, and a website has to appear in both periods to be included at all.

The preferred sample has 8,941 website-month observations from 425 websites. The six EU-domain groups are `.at`, `.be`, `.de`, `.eu`, `.fr`, and `.uk`. The four comparison groups are `.ch`, `.ru`, `.sx`, and `.tr`. Appendix Table 7 gives the website and site-month counts for each of these ten groups, and Appendix Table 8 reports pre-period means and standard deviations for each outcome in both groups. The imbalance here is substantial: German domains alone account for 361 of the 399 treated websites. That's a direct consequence of the German-heavy Cliqz contributor base, and it's the main reason I don't claim this estimates a representative EU-wide effect.

3.5 Parent-company concentration

WhoTracks.Me defines company reach as the proportion of observed page loads on which a parent company appears as a third party. Multiple companies can appear on the same page load, so raw reach values don't sum to one. For company i in month t, I construct

$$s_{it} = \frac{R_{it}}{\sum_j R_{jt}}, \tag{1}$$

where R_{it} is company reach in the global series. The shares in Equation 1 sum to one and describe each company's position relative to other observed tracking companies. They are not revenue shares, user shares,

or legal market shares.

I calculate the Herfindahl–Hirschman index,

$$HHI_t = 10{,}000 \sum_i s_{it}^2, \tag{2}$$

along with the combined shares of the four and ten largest companies, CR4 and CR10, and the effective number of companies, $10{,}000/HHI_t$. A higher HHI or CR4 means observed reach is more concentrated among a small set of parent companies.

I use `reach` rather than `site_reach` as the primary measure, since WhoTracks.Me changed the definition of site reach around February 2019. I also compare company-level concentration against concentration calculated across individual tracking services, which isolates the measurement effect of consolidating services under common ownership.

4 Empirical Framework

4.1 Website fixed-effects comparison

Let Y_{st} be an outcome for website s in month t. Let EU_s equal one when the website uses a supported EU-28 ccTLD, and let $Post_t$ equal one from June 2018 onward. The preferred model is

$$Y_{st} = \alpha_s + \lambda_t + \beta(EU_s \times Post_t) + \varepsilon_{st}, \tag{3}$$

where α_s are website fixed effects and λ_t are month fixed effects. The pre-period is May 2017–February 2018. March–May 2018 is omitted. The post-period is June 2018–May 2019.

Website fixed effects absorb stable differences between sites, including domain suffix, baseline popularity, and sustained tracking intensity. Month fixed effects absorb shocks that hit both samples at once, including broad changes in browsers, tracking technology, or the overall contributor count. The coefficient β is the differential post-period change of EU-domain websites relative to comparison-domain websites.

A causal reading needs more than the regression itself can offer. In the absence of GDPR, the two outcome series would need to follow similar paths, and the measurement transition would need to affect both groups in the same way. I check the first condition using raw trends and differential pre-period slopes. The second condition isn't something I can test directly. So I use "estimate" or "differential change" rather than "effect" when I refer to Equation 3.

4.2 Inference and sensitivity analysis

The treatment indicator varies at the ccTLD-group level, not independently across thousands of individual website-month rows. I cluster the standard errors for the preferred website-level model by ccTLD group and calendar month, which allows for arbitrary dependence among websites that share a suffix and among

observations that fall in the same month (Cameron et al., 2011). Since there are only ten supported ccTLD groups, I also aggregate to ccTLD-by-month means, give each group equal weight, cluster by group, and compute a restricted-residual wild-cluster bootstrap for the three prespecified primary outcomes (MacKinnon & Webb, 2018).

I report a differential linear slope over the ten pre-period months. A nonzero slope is evidence against a simple parallel-trends assumption (Bertrand et al., 2004). I don't treat an insignificant pre-period test as proof that trends are actually parallel, especially with only ten groups and a short pre-period to work with.

Sensitivity analyses extend the post-period to 24 months or the full available horizon, omit only May 2018 instead of the entire transition window, restrict to websites that are stable or fully balanced, weight by pre-period popularity, and replace common month effects with category-by-month effects. These specifications check whether the result depends on horizon, sample composition, or website category. They don't solve the problem posed by the contributor transition itself.

4.3 Descriptive concentration analysis

The concentration analysis is deliberately descriptive. I plot the monthly company HHI and compare the average over May 2017–February 2018 with the average over June 2018–May 2019. I also run segmented time-series regressions with a linear trend, a post-period level shift, and a post-period slope change, using Newey–West standard errors with three lags. These regressions summarize the recorded series. They are not used to assign causality, since the contributor transition overlaps with GDPR.

To identify the mechanical source of the HHI change, I average each company's raw reach within the pre- and post-periods, normalize the window averages, and decompose

$$\Delta HHI = 10{,}000 \sum_i \left(s^2_{i,post} - s^2_{i,pre} \right). \tag{4}$$

I then group the terms in Equation 4 by whether a company is observed in both windows, only before, or only after. This grouping is just an accounting identity. "Observed only after" is not the same thing as a company entering the market, and "observed only before" is not the same thing as a company exiting it.

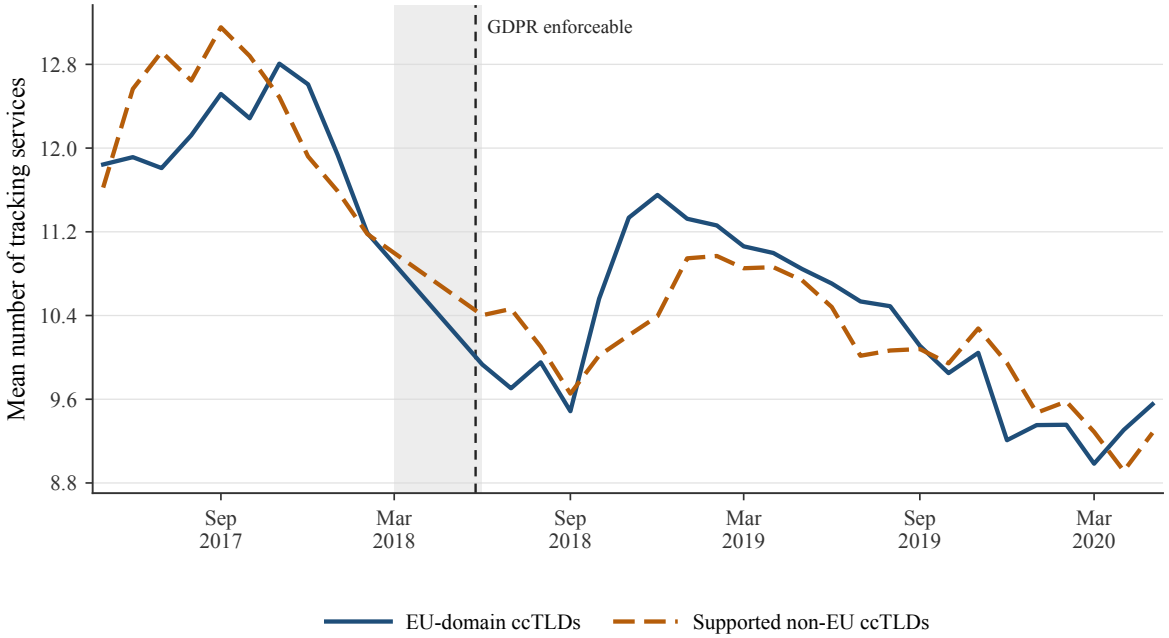
5 Results

5.1 Raw website trends

Figure 4 shows the mean number of tracking services on EU-domain and comparison-domain websites. The two samples aren't on parallel trajectories before GDPR. Comparison domains start out below EU domains in May 2017, exceed them by mid-2017, and converge again by early 2018. Both groups fall during the excluded transition period. EU domains exceed comparison websites during late 2018 and early 2019. The graph does not show a sharp policy break point on May 25, 2018.

The identifier-transmission trend series, shown in Appendix Figure 8, faces the same problem. The EU-

Figure 4: Monthly Mean Number of Recorded Tracking Services per Website



Notes: Monthly unweighted means for the websites in the supported ccTLD sample. The shaded interval, March–May 2018, is excluded from the preferred regressions. The dashed vertical line marks May 25, 2018. The plotted series go through May 2020 for context, but the preferred post-period goes through May 2019.

domain websites start below comparison websites and narrow the gap before the policy date. The regression results below quantify that convergence.

5.2 Preferred website estimates

Table 3 presents the preferred estimates. None of the five website fixed-effect coefficients is statistically distinguishable from zero. The number of tracking services increases by 1.051 per website ($SE = 0.946$), and the number of tracking companies increases by 1.004 ($SE = 0.822$). Any identifier transmission decreases by 0.010 and cookie-based tracking decreases by 0.007. Query-string identifier transmission moves the other way, increasing by 0.049.

The confidence intervals in this table give me more information than the significance labels alone. The confidence interval for the number of tracking services is $[-0.80, 2.90]$, against a pre-period mean of 12.11 for EU-domain websites. The lower bound corresponds to about a 6.6 percent decline. So the interval can't rule out a modest decline, but it does rule out a decline anywhere close to one quarter of the pre-period mean under this specification. The confidence interval for any identifier transmission ranges between -0.047 and 0.027: a decrease of 4.7 percentage points on one end, an increase of 2.7 points on the other.

The country-month estimates are less precise than the website models, but they point in the same direction. The wild-cluster bootstrap p-values for tracking services, tracking companies, and any identifier transmission are 0.530, 0.521, and 0.708 respectively. This shows me that millions of underlying records do not necessarily imply precise policy inference once I've clustered down to just ten ccTLD groups.

Table 3: Preferred Website-Level Estimates of Differential Tracking Changes

	Tracking services	Tracking companies	Any identifier transmission	Cookie-based tracking	Query-string identifier
Panel A. Estimated differential post-period changes					
Website fixed-effects estimate	1.051	1.004	-0.010	-0.007	0.049
	(0.946)	(0.822)	(0.019)	(0.018)	(0.048)
Equal-weight country-month estimate	1.376	1.556	-0.019	-0.002	0.054
	(2.226)	(1.848)	(0.049)	(0.040)	(0.078)
Wild-cluster bootstrap p-value	0.530	0.521	0.708	–	–
Differential pre-period slope	0.123*	0.087	0.0068***	0.0073***	0.0090*
	(0.072)	(0.069)	(0.0014)	(0.0012)	(0.0048)
Pre-period EU-domain mean	12.106	9.103	0.872	0.856	0.428

	Main website fixed-effects model	Equal-weight country-month model	Pre-period slope model
Panel B. Specification details			
Unit of observation	Website-month	ccTLD-month mean	Website-month
Website fixed effects	Yes	No	Yes
ccTLD fixed effects	No	Yes	No
Calendar-month fixed effects	Yes	Yes	Yes
Post-period indicator estimated	Yes	Yes	No
Equal weight across ccTLD groups	No	Yes	No
Pre-period observations only	No	No	Yes
Covariance estimator	ccTLD × month	ccTLD	ccTLD × month
Wild-cluster bootstrap reported	No	Yes[a]	No
Observations	8,941	220	3,542
Independent units	425 websites	10 ccTLD groups	425 websites

Notes: The sample uses the global WhoTracks.Me series and compares websites using EU-28 country-code top-level domains (ccTLDs) with websites using supported non-EU country-code domains. The pre-period is May 2017–February 2018, March–May 2018 is omitted, and the post-period is June 2018–May 2019. Panel A reports the estimated differential change for each outcome. The main website model includes website and calendar-month fixed effects. The equal-weight country-month model collapses the data to ccTLD-by-month means, gives each ccTLD group equal weight, and includes ccTLD and calendar-month fixed effects. The differential pre-period slope is estimated only within the ten pre-period months, using website and calendar-month fixed effects.

[a] Wild-cluster bootstrap p-values are reported for the three prespecified primary outcomes from the equal-weight country-month model, with 999 replications.

Standard errors are in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

The pre-period slope is a substantive issue. Any identifier transmission increases 0.0068 more per month for EU-domain websites compared to comparison websites, and cookie-based tracking increases 0.0073 more per month for EU domains. These estimates are highly statistically significant. The tracking-service count also has a positive differential slope of 0.123 per month, though with a p-value of 0.088. These patterns are consistent with pre-GDPR convergence, and they weaken the identifying assumption behind Equation 3.

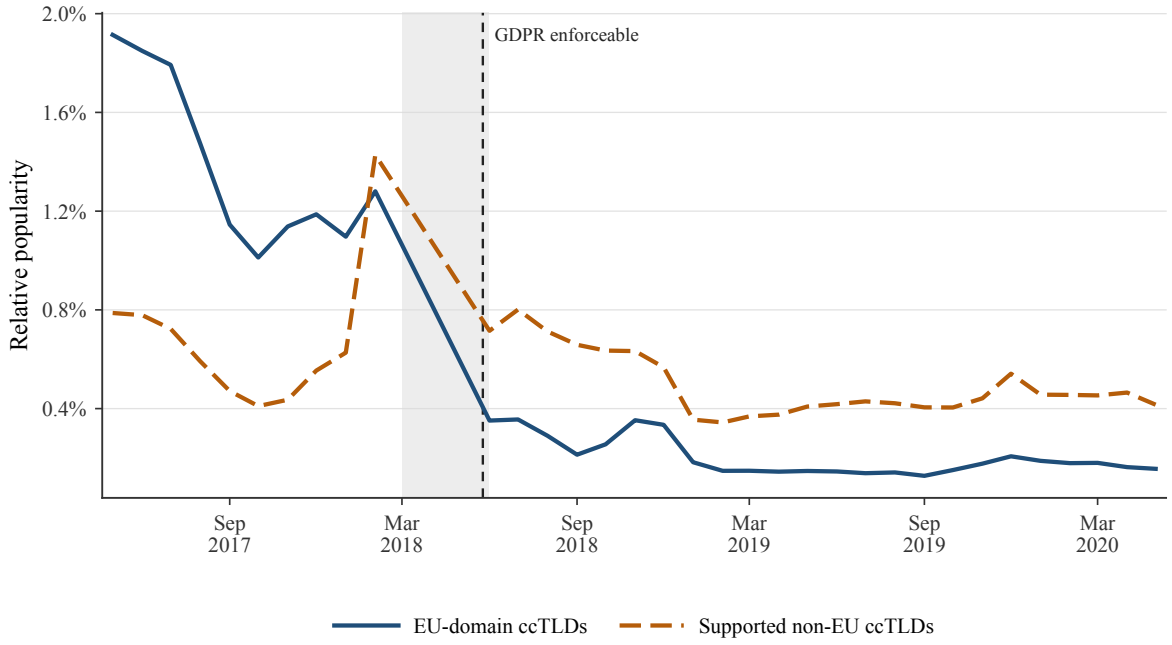
5.3 Measurement and composition diagnostics

Website outcomes are not the only variables changing differentially. Relative popularity decreases by 0.0098 for EU-domain websites after the transition, with a standard error of 0.0030. This is large compared with the pre-period mean of 0.0137 among EU-domain websites. Relative popularity here is normalized to the most popular website in the monthly file, so this estimate should not be interpreted as a 71 percent decrease in traffic. What it tells me is that the position of these websites within the contributor-generated traffic distribution has shifted differently between the two groups. Figure 5 shows this trend and illustrates that the composition of websites represented in the two groups changes at the same time as the policy transition.

Blocking is a second caution signal, though not a good control. `has_blocking` was added to WhoTracks.Me in December 2017, and the series has its own discontinuities. A discontinuity in blocking could affect recorded tracker counts independently of any actual change in websites' third-party integrations. Because this measure is not stable over the full period, I use it as a diagnostic rather than as a regression control or a negative-control outcome.

Figure 5 and Figure 3 taken together help explain why fixed effects are necessary but not sufficient. Month fixed effects soak up a global, platform-wide change affecting both EU and comparison domains. They do not soak up a change in which types of websites or contributors become represented in each group.

Figure 5: Monthly Relative Popularity in the Preferred Website Sample



Notes: Monthly unweighted means of WhoTracks.Me's relative popularity measure, expressed on the vertical axis as a percentage of the most popular website in the same file. Popularity is not an absolute traffic count. The shaded interval is excluded from the preferred regressions, and the dashed vertical line marks May 25, 2018. The series is a composition diagnostic, not a welfare or demand measure.

5.4 Sensitivity across specifications

Appendix Table 9 demonstrates that the estimate is not hiding somewhere else in the specification space. The 24-month estimates continue to be positive for service and company counts, and negative but small for identifier transmission. The stable and balanced samples generate similar results. Weighting by baseline popularity reverses the sign of the count estimates, but these coefficients remain imprecise either way. Category-by-month fixed effects do not uncover a large negative shift in any of the outcomes.

The full post-period estimate for any identifier transmission is −0.100, with a standard error of 0.053. This horizon runs through February 2025, almost seven years after GDPR. The estimate absorbs a lot of later changes to browsers, websites, consent mechanisms, TrackerDB, and the contributor population itself. It should not be taken as an estimate of the immediate effect of the 2018 regulation. Lack of consistency across outcomes and horizons is itself evidence against one single large break caused by GDPR.

Appendix Table 10 drops each of the ten ccTLD groups separately from the equal-weight country-month model. None of these omissions produces a precise negative estimate of tracker count. The tracker-count coefficients are sensitive to the choice of comparison domain: dropping `.tr` reduces the tracker estimate from +1.38 to −0.54, while dropping `.ru` increases it to +3.06. The any-tracking coefficient flips sign when either `.eu` or `.ru` is dropped. This is what I'd expect from a design based on four comparison groups, not evidence of an invalid domain. If anything, this exercise reinforces why I should report the effective number of clusters, and not treat 8,941 website-month observations as 8,941 independent policy comparisons.

The discrepancy between my estimates and the decline reported by Miller et al. (2024) is best understood as a comparison of designs, not a contest of statistical significance. Their balanced panel of publishers, treatment assignment, and standard-error construction target a different comparison than mine. My preferred specification sacrifices coverage in order to avoid confounding regional file creation with treatment, and it reports the resulting ten-group design truthfully.

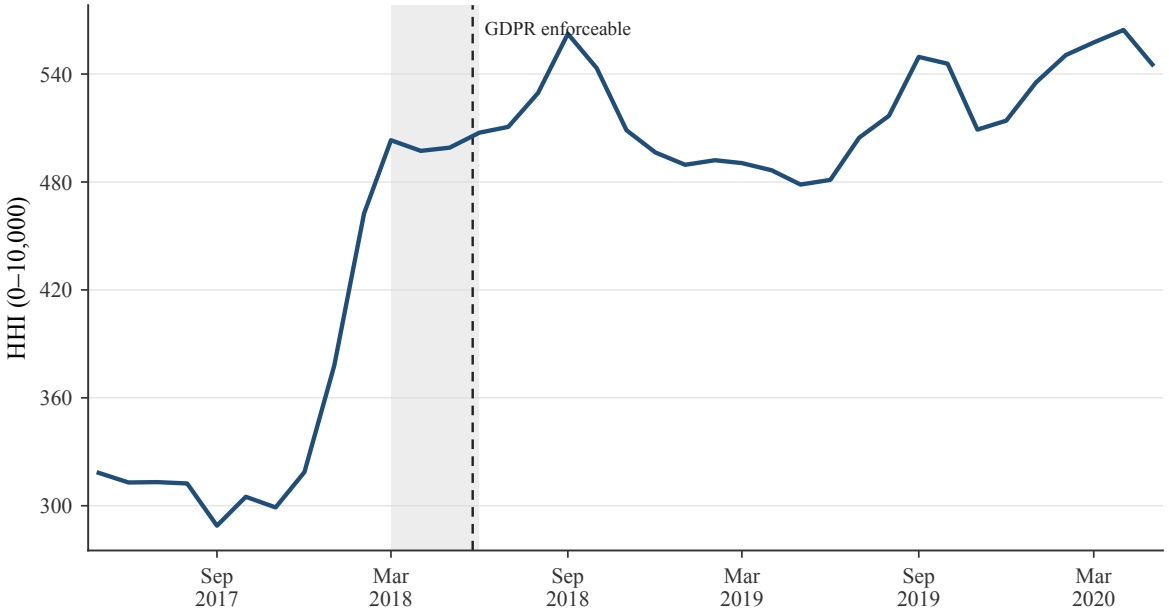
My conclusion from the website analysis is not that GDPR had exactly zero effect. The design here is not powerful enough to make such a claim. The conclusion is narrower: the WhoTracks.Me comparison does not provide credible evidence for the substantial privacy improvement suggested by the pooled hybrid-region specification.

5.5 Parent-company concentration

Figure 6 plots HHI based on normalized parent-company reach in the global series. Concentration starts rising well before GDPR becomes enforceable. HHI stands at 318.7 in December 2017, 377.9 in January 2018, and 462.5 in February. The series passes 500 in March, the same month the Ghostery contributor transition starts. It stays around 500 for most of the following two years.

Table 4 summarizes the change. Average monthly HHI rises from 330.88 in the pre-period to 507.97 in the post-period. CR4 rises by 5.7 percentage points, CR10 by 5.2 points, and the effective number of companies falls from 30.76 to 19.73. At the same time, the number of companies appearing in the monthly file rises by

Figure 6: Monthly Parent-Company HHI in the Global WhoTracks.Me Series



Notes: Monthly HHI calculated from parent-company reach in the global WhoTracks.Me series. Reach of each company is normalized to total observed reach of companies in the same month. The shaded interval marks March–May 2018; the dashed vertical line marks May 25, 2018. The series is descriptive and does not isolate GDPR from the contributor transition.

239, and total company reach falls by 23 percent.

Table 4: How Parent-Company Concentration Changed Around GDPR (Descriptive)

Measure	Pre-period	Post-period	Change	Percent change
HHI (0–10,000)	330.88	507.97	+177.10	+53.5%
Top-four share	27.5%	33.2%	+5.7 pp	+20.8%
Top-ten share	38.4%	43.6%	+5.2 pp	+13.4%
Leading-company share	14.9%	20.0%	+5.1 pp	+34.0%
Observed companies	575.40	814.50	+239.10	+41.6%
Total company reach	5.32	4.09	-1.23	-23.1%
Effective number of companies	30.76	19.73	-11.03	-35.9%

Notes: For each month, a company's reach is divided by total observed company reach in the global series that month. The pre-period is May 2017–February 2018, the post-period is June 2018–May 2019, and March–May 2018 is omitted. Because several companies can appear on the same page load, these normalized reach shares describe relative tracking presence rather than revenue or sales shares. The table averages monthly measures within each window; it is descriptive and is not a causal estimate of any GDPR effect.

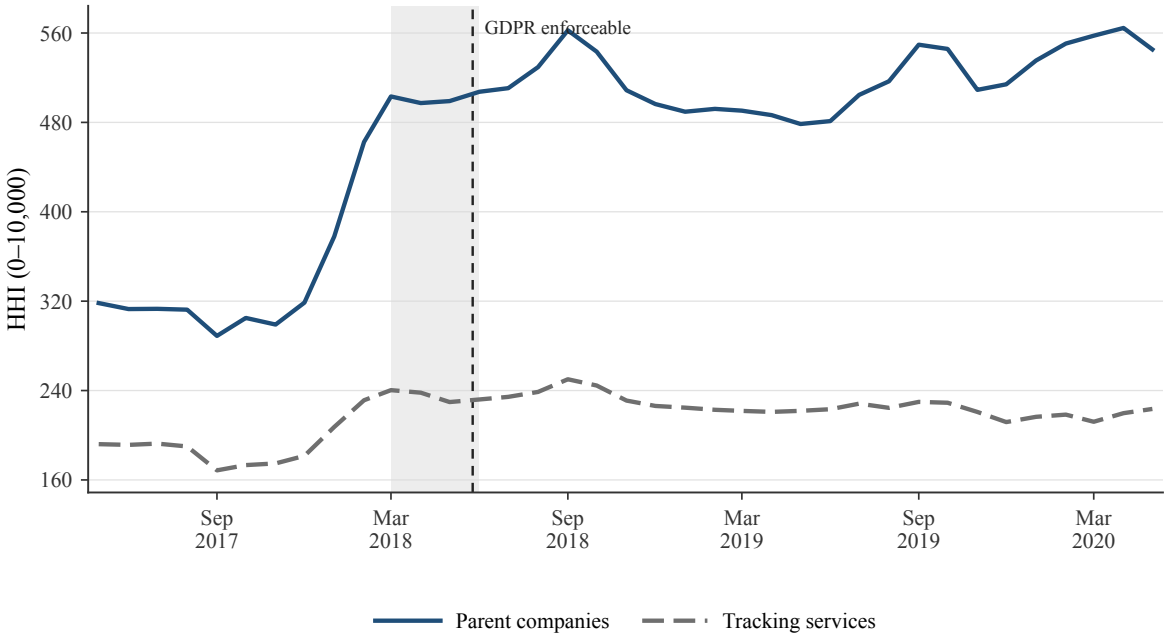
An increase in the number of observed companies and a rise in concentration are not actually in contradiction. New or newly identified companies can enter the long tail while a leading firm continues to gain relative reach. The decline in total reach matters here too. Google's average raw reach changes little, from 0.786 to 0.816, but total reach declines enough that Google's normalized reach share rises substantially. What's happening here is a change in relative presence, not an equal rise in every measure of Google's activity.

Appendix Table 12 checks this using segmented regressions instead of simple period averages. HHI jumps by 112.2 points at the post-period break (SE = 65.6), and its slope after the break is 15.3 points per month lower than its pre-period slope (SE = 4.9). Most of the change shows up as a one-time shift, not as an accelerating trend. CR4, CR10, and the leading company's share follow a similar pattern: a positive shift at the break, alongside a flatter or falling slope afterward. These regressions summarize the same descriptive series a different way. They do not identify a causal GDPR effect any more than the period averages in Table 4 do.

5.6 Why ownership aggregation matters

Figure 7 plots company-level HHI against tracking-service HHI. Parent-company HHI is higher than tracking-service HHI by an average of 249.8 points in the global series, and company CR4 is higher than service CR4 by 9.1 percentage points. Not aggregating services under their parent companies systematically understates concentration.

Figure 7: Tracking-Service and Parent-Company HHI in the Global Series



Notes: Both series use reach normalized within the global monthly file. The service series treats each tracking service as a separate entity. The company series aggregates services under their historical parent-company identifier. The difference measures ownership aggregation, not a causal policy response.

This changes the economic interpretation. Service diversity can stay high even when ownership concentration is high. And one company can increase its relative reach through several tracking services at once. When the question is market structure, the company-level HHI is the right measure to report; the tracking-service HHI works well as an additional descriptive statistic.

5.7 Decomposing the HHI increase

Table 5 decomposes the increase in HHI using Equation 4. HHI calculated from window-average shares increases by 181.81 points. Companies observed in both windows account for 180.32 of those points, or 99.2 percent of the total increase. Companies observed only after the transition account for 1.56 HHI points, and companies observed only before the transition account for −0.06 HHI points.

Table 5: Exact Decomposition of the Rise in Parent-Company HHI

Company group	Number	Pre-period share	HHI contribution
Observed in both windows	667	99.3%	+180.32
Observed only after	245	0.0%	+1.56
Observed only before	22	0.7%	-0.06
Total change	1,047	100.0%	+181.81

Notes: This calculation averages each company's raw reach separately within the pre- and post-periods, then normalizes those window averages to sum to one. HHI is 323.06 before and 504.87 after. This differs slightly from the 330.88-to-507.97 change reported in Table 4, because HHI is nonlinear: this table computes HHI once from window-average shares, while Table 4 averages the monthly HHI values themselves. "Observed only after" and "observed only before" describe appearance in the published files, not verified economic entry or exit.

This rules out the initial conjecture that the rise in concentration was driven by the disappearance of small trackers. The data don't tell me why the shares of continuing firms changed in the first place: market evolution, shifts in website or user composition, blocking, reclassification of services, and GDPR itself are all still possible explanations. The immediate effect, however, is clear: almost all of the increase comes from reallocating shares among firms that were present in the files the whole time.

Google is driving this reallocation. Its normalized reach share increases from 14.78 to 19.94 percent, and the change in its squared share alone contributes 179.10 HHI points, or 98.5 percent of the total exact increase. Other firms show offsetting changes in relative share: Facebook, Microsoft, Comscore, Twitter, and Cloudflare gain relative share, while INFOnline, AppNexus, Criteo, and several other advertising firms lose share. Appendix Table 11 lists the leading firms.

The title of this thesis asks who wins and loses. As a purely descriptive statement, Google gains relative reach in these files. This does not mean it has gained because of GDPR. The increase starts before GDPR becomes enforceable and coincides with the contributor transition. That's an important distinction here, and I don't think it's one this design can collapse.

6 Discussion

The new evidence does not indicate a causal trade-off between privacy and competition. The preferred website estimates do not reveal a differential reduction in tracking after GDPR, and the direct prevalence measures show pre-trends that are inconsistent with a simple difference-in-differences interpretation. Concentration

among parent companies increases, but the increase starts before GDPR becomes enforceable. These two facts cannot be read together as evidence that GDPR improved privacy while concentrating the market.

The lack of causal evidence does not necessarily mean that GDPR had no effect. The regulation altered legal obligations, consent procedures, data policies, investments, and product development decisions in various industries, as documented in other papers (Aridor et al., 2023; Janßen et al., 2022; Jia et al., 2021; Johnson et al., 2023; Peukert et al., 2022). It means that this particular data source does not separate GDPR from a simultaneous change in measurement well enough to estimate the broad causal effect claimed in the original analysis.

Three lessons generalize beyond GDPR. First, data provenance is part of identification. The panel generated by a platform can change because the platform alters who contributes, what gets blocked or classified, or which entities are published at all. Month fixed effects remove shocks common to both groups, but they do not solve differential changes in panel composition across treatment groups.

Second, uncertainty should match the source of policy variation. Clustering standard errors at the website level would treat thousands of sites as though they were independent sources of GDPR variation, even though exposure to the regulation is assigned by a handful of domain groups. The much larger standard errors that come from group-level inference are not a flaw in the analysis; they are a more accurate description of what the comparison can actually identify.

Third, the unit of market structure must match the economic question. Tracker services are not firms. Parent-company aggregation substantially raises measured concentration and changes the identity of the leading entities. Normalized reach shares are still imperfect: they measure presence, not prices, revenue, profits, or welfare. Even so, they come closer to ownership concentration than service-level shares do.

The analysis has a number of limitations. A ccTLD is an imperfect proxy for legal exposure. The preferred sample is small, weighted heavily toward Germany, and not representative of the EU as a whole. WhoTracks.Me does not cover every page or every tracking method, and pages with no detected trackers go unreported in the first place. Historical company mappings and publication thresholds can also change over time. I cannot observe user consent, welfare, compliance expenditure, or firm revenue. These limits rule out welfare claims and keep me from separating regulatory compliance from other changes happening in the tracking ecosystem.

A more convincing causal study would need an independently sampled and consistently measured panel spanning both sides of the policy date, a treatment definition based on actual user or firm legal exposure, and enough geographic units to support credible inference. Repeated web crawls of a fixed set of websites could help here, especially if paired with historical company mappings and a transparent record of any measurement changes. Such a design would complement the behavioral data from browser users, rather than simply adding more observations drawn from the same shifting source.

7 Conclusion

This paper reassesses the evidence on GDPR and online tracking using public WhoTracks.Me data. The data are rich, but the contributor base, geographic coverage, and some measurement definitions all change right around GDPR implementation. Once I restrict the website comparison to a consistent global series, define groups using ccTLDs, include website and month fixed effects, and conduct inference at the level of the domain-group comparison, I do not find robust evidence of a large post-GDPR decline in recorded tracking.

I also find that concentration among parent companies in observed tracking reach starts rising around early 2018. The rise happens before GDPR takes effect and therefore cannot be attributed to the regulation using this data series alone. The exact HHI accounting shows that the increase is driven almost entirely by redistribution among continuing companies, especially a rise in Google's relative reach, rather than by the disappearance of small companies. Aggregating tracking services up to their parent companies also reveals substantially higher concentration than a service-level calculation would show.

The main contribution of this paper is a boundary on what this data can actually establish. Large digital trace datasets can produce estimates that look precise, even while the effective policy comparison underneath remains small and unstable. A credible analysis needs clear units of observation, stable measurement, inference set at the level of treatment, and conclusions that follow from the diagnostics rather than around them. Under those requirements, a simple story of privacy gains and competitive losses gives way to something more careful: ownership concentration rose in the recorded data, but the causal effect of GDPR itself remains unresolved.

A Supplementary Data, Definitions, and Sample Construction

Table 6: Definitions of the Main Variables and Measures

Term	Definition
Website	A first-party domain appearing among the popular sites reported in a monthly WhoTracks.Me file. The set of published websites changes over time.
Country-code top-level domain (ccTLD)	The country or territory suffix at the end of a domain name, such as `.de` for Germany or `.fr` for France. A ccTLD is a proxy for a website's geographic orientation; it does not establish the owner's legal location or the user's location.
Tracking service	A group of third-party domains that WhoTracks.Me and TrackerDB attribute to the same tracking technology or service.
Parent company	The organization to which TrackerDB assigns one or more tracking services. The official `companies.csv` files aggregate services at this level.
Tracking services per website	Average number of classified tracking services detected on the website's observed page loads.
Tracking companies per website	Average number of classified parent companies detected on the website's observed page loads.
Cookie-based tracking	Proportion of observed page loads on which the browser sent a cookie or the third party returned a `Set-Cookie` header.
Query-string identifier transmission	Proportion of observed page loads on which WhoTracks.Me detected an identifier-like value in a request's query string. The original variable name is `bad_qs`; it is not a general measure of browser fingerprinting.
Any identifier transmission	Proportion of observed page loads on which WhoTracks.Me detected identifier transmission through either cookies or query-string parameters. The original variable name is `tracked`.
Company reach	Proportion of observed page loads on which a parent company appears as a third party. Several companies can appear on one page load, so raw company reach does not sum to one.
Normalized company share	Company reach divided by the sum of company reach in the same month and region. It measures relative observed tracking presence, not revenue or sales share.
HHI	Sum of squared normalized company shares, multiplied by 10,000.

Term	Definition
CR4 and CR10	Sum of the four or ten largest normalized company shares in a month.
Effective number of companies	Reciprocal of HHI on the zero-to-one scale; equivalently, $10{,}000/HHI$ when HHI is reported on the zero-to-10,000 scale used elsewhere in this paper.

Table 7: The Ten Supported ccTLD Groups in the Preferred Website Sample

Domain suffix	Group	Websites	Total site-months	Pre-period site-months	Post-period site-months
`.de`	EU-domain	361	7,633	3,346	4,287
`.at`	EU-domain	17	354	155	199
`.fr`	EU-domain	13	282	126	156
`.uk`	EU-domain	5	98	38	60
`.eu`	EU-domain	2	44	20	24
`.be`	EU-domain	1	22	10	12
`.ch`	Comparison	13	258	109	149
`.ru`	Comparison	8	144	48	96
`.tr`	Comparison	3	65	29	36
`.sx`	Comparison	2	41	17	24
Total	–	425	8,941	3,898	5,043

Notes: The United Kingdom is in the EU-domain group because it was an EU member when GDPR became enforceable. The regional `.eu` suffix is also included. A group is supported when it has at least five pre-period months, six post-period months, ten pre-period site-months, and ten post-period site-months. Websites must appear in both periods. German domains account for 361 of the 399 EU-domain websites.

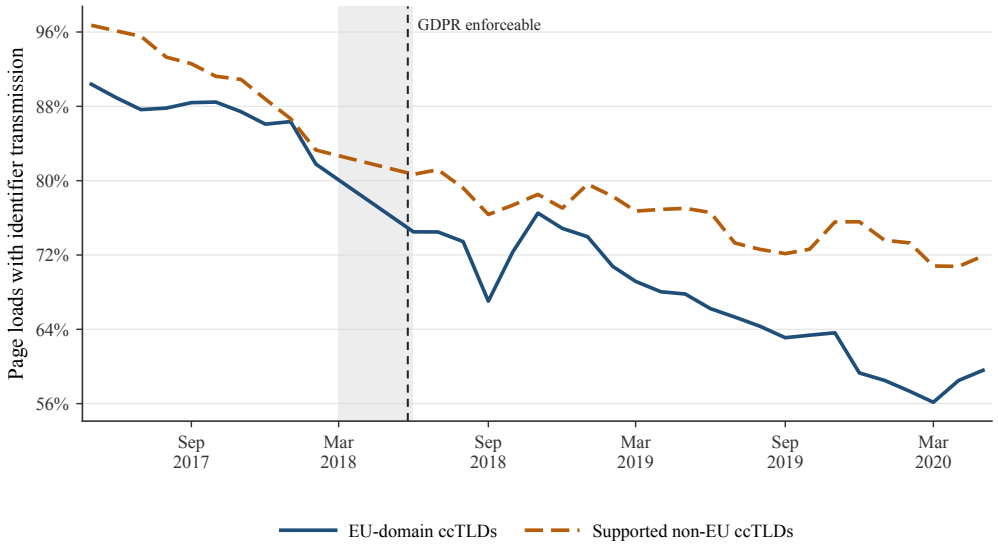
Table 8: Pre-Period Summary Statistics for the Preferred Website Sample

	EU-domain websites		Comparison websites	
Outcome	Mean	SD	Mean	SD
Tracking services	12.106	7.673	12.254	6.470
Tracking companies	9.103	6.115	9.489	4.861
Any identifier transmission	0.872	0.157	0.910	0.107
Cookie-based tracking	0.856	0.183	0.904	0.109
Query-string identifier transmission	0.428	0.197	0.414	0.237

Notes: The pre-period is May 2017–February 2018. These statistics are unweighted website-month summaries. The prevalence outcomes are proportions between zero and one.

B Supplementary Results and Diagnostics

Figure 8: Monthly Identifier Transmission in the Preferred Website Sample



Notes: Monthly unweighted means for the supported ccTLD sample. The vertical axis reports the percentage of observed page loads on which WhoTracks.Me detects identifier transmission through cookies or query-string parameters. As in Figure 4, the shaded band covers the excluded March–May 2018 window, and the dashed line marks May 25, 2018.

Table 9: Sensitivity Analyses for the Preferred Website Estimates

Specification	Tracking services	Tracking companies	Any identifier transmission
Preferred specification (12-month post-period)	1.051	1.004	-0.010
	(0.946)	(0.822)	(0.019)
24-month post-period	1.007	1.087	-0.027
	(0.945)	(0.708)	(0.018)
Full post-period	-0.595	0.005	-0.100*
	(1.110)	(0.771)	(0.053)
Exclude May 2018 only	0.669	0.761	-0.013
	(0.690)	(0.559)	(0.016)
Stable websites	1.486	1.368	0.001
	(1.024)	(0.933)	(0.014)
Balanced websites	1.056	0.963	-0.014
	(0.757)	(0.634)	(0.011)
Baseline-popularity weighted	-0.457	-0.385	-0.012
	(0.464)	(0.349)	(0.025)
Category-by-month fixed effects	0.837	0.797	-0.005
	(0.872)	(0.764)	(0.024)
Log(1+outcome)	0.090	0.105	–
	(0.055)	(0.067)	

Notes: Each coefficient is the EU-domain-by-post estimate. Except where the row label indicates otherwise, regressions include website and month fixed effects and cluster standard errors by top-level-domain group and month. The full post-period extends through February 2025 and should be read as a long-run association, not an immediate GDPR effect. The log specification applies only to the count outcomes.

Standard errors are in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 10: Leave-One-ccTLD-Out Estimates from the Equal-Weight Country-Month Model

ccTLD omitted	Tracking services	Tracking companies	Any identifier transmission
None (baseline; all ten groups)	1.376 (2.226)	1.556 (1.848)	-0.019 (0.049)
.at (Austria)	1.437 (2.269)	1.618 (1.883)	-0.022 (0.057)
.be (Belgium)	1.282 (2.267)	1.449 (1.880)	-0.016 (0.057)
.de (Germany)	1.577 (2.256)	1.722 (1.873)	-0.020 (0.057)
.eu (European Union)	1.583 (2.255)	1.731 (1.871)	0.015 (0.036)
.fr (France)	1.254 (2.265)	1.421 (1.877)	-0.038 (0.051)
.uk (United Kingdom)	1.126 (2.248)	1.393 (1.873)	-0.035 (0.053)
.ch (Switzerland)	1.325 (2.984)	1.617 (2.477)	-0.022 (0.056)
.ru (Russia)	3.061 (1.972)	2.883 (1.739)	0.006 (0.045)
.sx (Sint Maarten)	1.659 (2.962)	1.814 (2.454)	-0.020 (0.056)
.tr (Turkey)	-0.540 (1.553)	-0.092 (1.158)	-0.040 (0.048)

Notes: Each cell reports the coefficient on EU-domain × post and its ccTLD-clustered standard error from an equal-weight ccTLD-by-month regression with ccTLD and month fixed effects. The first row is the full ten-group sample. Each subsequent row drops one ccTLD: the six EU-domain rows (`.at` through `.uk`) are treated groups, and the four remaining rows (`.ch`, `.ru`, `.sx`, `.tr`) are comparison groups. The exercise is a sensitivity diagnostic, not a correction for the small number or nonrandom composition of the groups.

Table 11: Leading Parent Companies by Normalized Tracking Reach

Company	Pre-period share	Post-period share	Share change	Rank change
Google	14.78%	19.94%	+5.16 pp	0
Facebook	5.77%	6.82%	+1.05 pp	0
Amazon Associates	3.49%	4.15%	+0.66 pp	0
INFOnline	3.29%	0.75%	-2.54 pp	-14
Criteo	2.52%	1.61%	-0.92 pp	-3
Comscore	1.30%	2.24%	+0.94 pp	+8
AppNexus	2.20%	1.19%	-1.01 pp	-7
Twitter	1.22%	2.15%	+0.93 pp	+9
Microsoft	0.33%	1.93%	+1.60 pp	+53
Cloudflare	1.02%	1.72%	+0.70 pp	+9

Notes: Shares normalize average company reach by the sum of average company reach within the corresponding window. A positive rank change indicates movement toward a higher rank. Names are lightly formatted versions of the historical WhoTracks.Me company identifiers.

Table 12: Descriptive Segmented Time-Series Estimates for Global Company Concentration

	Post-period level shift		Post-period slope change	
Measure	Estimate	SE	Estimate	SE
HHI	112.244*	65.551	-15.303***	4.924
CR4	3.868	2.369	-0.412**	0.198
CR10	4.286*	2.404	-0.216	0.194
Leading-company share	2.429	1.718	-0.501***	0.135

Notes: Regressions use the global company-reach series, omit March–May 2018, and include a linear time trend, a post-period indicator, and a post-period trend. Standard errors are Newey–West with three lags. Share outcomes are reported in percentage points. These models summarize the recorded series; they are not causal estimates because the contributor transition overlaps GDPR implementation.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

References

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, *54*(2), 442–492. https://doi.org/10.1257/jel.54.2.442

Aridor, G., Che, Y.-K., & Salz, T. (2023). The effect of privacy regulation on the data industry: Empirical evidence from the GDPR. *RAND Journal of Economics*, *54*(4), 695–730. https://doi.org/10.1111/1756-2171.12455

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, *119*(1), 249–275. https://doi.org/10.1162/003355304772839588

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249. https://doi.org/10.1198/jbes.2010.07136

Campbell, J., Goldfarb, A., & Tucker, C. (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy*, *24*(1), 47–73. https://doi.org/10.1111/jems.12079

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

GDPR Enforcement Tracker. (2025). Fines database. https://enforcementtracker.com/

Godinho de Matos, M., & Adjerid, I. (2022). Consumer consent and firm targeting after GDPR: The case of a large telecom provider. *Management Science*, *68*(5), 3330–3378. https://doi.org/10.1287/mnsc.2021.4054

Goldfarb, A., & Tucker, C. E. (2011). Privacy regulation and online advertising. *Management Science*, *57*(1), 57–71. https://doi.org/10.1287/mnsc.1100.1246

Hagiu, A., & Wright, J. (2023). Data-enabled learning, network effects, and competitive advantage. *RAND Journal of Economics*, *54*(4), 638–667. https://doi.org/10.1111/1756-2171.12453

Janßen, R., Kesler, R., Kummer, M. E., & Waldfogel, J. (2022). GDPR and the lost generation of innovative apps. *Science*, *376*(6592), 484–488. https://doi.org/10.1126/science.abn1479

Jia, J., Jin, G. Z., & Wagman, L. (2021). The short-run effects of the General Data Protection Regulation on technology venture investment. *Marketing Science*, *40*(4), 661–684. https://doi.org/10.1287/mksc.2020.1271

Johnson, G. A., Shriver, S. K., & Goldberg, S. G. (2023). Privacy and market concentration: Intended and unintended consequences of the GDPR. *Management Science*, *69*(10), 5695–5721. https://doi.org/10.1287/mnsc.2023.4709

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, *110*(9), 2819–2858. https://doi.org/10.1257/aer.20191330

Karaj, A., Macbeth, S., Berson, R., & Pujol, J. M. (2018). WhoTracks.Me: Shedding light on the opaque world of online tracking.

MacKinnon, J. G., & Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *Econometrics Journal*, *21*(2), 114–135. https://doi.org/10.1111/ectj.12107

Miller, K. M., Lukic, K., & Skiera, B. (2024). The impact of the General Data Protection Regulation (GDPR) on online tracking. https://arxiv.org/abs/2411.06862

Natural Earth. (2025). Admin 0 – countries, 1:110m cultural vectors. https://www.naturalearthdata.com/downloads/110m-cultural-vectors/110m-admin-0-countries/

Peukert, C., Bechtold, S., Batikas, M., & Kretschmer, T. (2022). Regulatory spillovers and data governance: Evidence from the GDPR. *Marketing Science*, *41*(4), 746–768. https://doi.org/10.1287/mksc.2021.1339

Prüfer, J., & Schottmüller, C. (2021). Competing with big data. *Journal of Industrial Economics*, *69*(4), 967–1008. https://doi.org/10.1111/joie.12259

WhoTracks.Me. (n.d.). WhoTracks.Me data documentation. https://github.com/whotracksme/whotracks.me

Yu, Z., Macbeth, S., Modi, K., & Pujol, J. M. (2016). Tracking the trackers. *Proceedings of the 25th International Conference on World Wide Web*, 121–132. https://doi.org/10.1145/2872427.2883028