“Please Respect Our Terms and Conditions”: A Causal Analysis of GDPR Impact on Privacy Policies

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37368591

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
“Please Respect Our Terms and Conditions”: A Causal Analysis of GDPR Impact on Privacy Policies

Natalie Margulies

Presented to the Department of Computer Science and Department of Statistics
in partial fulfillment of the requirements for a Bachelor of Arts

Harvard University
March 26, 2021
Abstract

The General Data Protection Regulation (GDPR) has been widely praised as the most consequential privacy law in history. However, GDPR causal effects have never been formally analyzed, and all GDPR praises are largely unsubstantiated. This thesis constructs a database of 317,396 privacy policies to overcome previous data limitations and formally quantifies the causal impact of GDPR regulation on privacy policies.

The thesis begins by addressing the open question of whether GDPR regulation has significantly changed privacy policies. I find that the GDPR has substantially changed privacy policies since its adoption: GDPR websites, on average, have changed their privacy policies 12.45% more and updated their privacy policies 25.43% more frequently than a non-GDPR control, ceteris paribus. The thesis next addresses the more nuanced question of how GDPR regulation has changed privacy policies. I identify a clear tension between the GDPR mandate for “concise and readable” privacy policies and additional GDPR Article disclosure requirements: the GDPR has made privacy policies more compliant with various GDPR articles but also less accessible.

Questions of GDPR efficacy are becoming particularly relevant to policy discussions. Many countries including New Zealand, India, South Africa, and the United States are modeling national privacy developments off of GDPR legislature. Further research on the nuanced effects of GDPR regulation on privacy policies is necessary for the GDPR to serve as a global paragon of privacy law successfully.
Acknowledgments

Thank you to my wonderful thesis advisors, Jim Waldo and Joe Blitzstein, who wholly inspired my interest in privacy.

Thank you to Susan Murphy, Jose Zubizarreta, Kosuke Imai, and Nikita Samarin, who all offered incredibly helpful feedback and advice.

Lastly, thank you to friends, family, and blockmates, who have all listened to my privacy spiel too many times to count.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Related Work</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Cross-Sectional</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Longitudinal</td>
<td>6</td>
</tr>
<tr>
<td>3 Methods</td>
<td>8</td>
</tr>
<tr>
<td>3.1 Basic Notation</td>
<td>8</td>
</tr>
<tr>
<td>3.2 Assumptions</td>
<td>10</td>
</tr>
<tr>
<td>3.3 Difference-in-Difference Estimator</td>
<td>13</td>
</tr>
<tr>
<td>3.4 Survival Analysis</td>
<td>14</td>
</tr>
<tr>
<td>4 Database</td>
<td>17</td>
</tr>
<tr>
<td>4.1 Website Dataset</td>
<td>17</td>
</tr>
<tr>
<td>4.2 Privacy Policy Dataset</td>
<td>19</td>
</tr>
<tr>
<td>4.3 GDPR Classification</td>
<td>22</td>
</tr>
<tr>
<td>4.4 Matching</td>
<td>25</td>
</tr>
<tr>
<td>4.5 Limitations</td>
<td>27</td>
</tr>
<tr>
<td>5 Quantifying GDPR Impact</td>
<td>28</td>
</tr>
<tr>
<td>5.1 Overall</td>
<td>28</td>
</tr>
<tr>
<td>5.2 Survival Analysis</td>
<td>29</td>
</tr>
<tr>
<td>5.3 Update Frequency</td>
<td>32</td>
</tr>
</tbody>
</table>
List of Figures

4.1 **GDPR Classifier Confusion Matrix** ............................................. 24

5.1 **Average Period-by-Period Privacy Policy Change for Matched GDPR and non-GDPR Websites** .................................................. 29

5.2 **Histogram of Time until First Update: Distributions of \( T_1 \) for Matched GDPR and non-GDPR Websites** ........................................ 31

5.3 **Estimated Survival and Hazard \( (T_1)\): Non-Parametric Estimates for Matched GDPR and non-GDPR Websites Survival and Hazard** .................................. 31

5.4 **Time until Second and Third Update Histogram: Distribution of \( T_2 \) and \( T_3 \) for Matched GDPR and non-GDPR Websites** ........................................ 32

5.5 **Estimated Survival and Hazard \( (T_2 \text{ and } T_3)\): Non-Parametric Estimates for Matched GDPR and non-GDPR Websites Survival** ........................................ 32

5.6 **Period-by-Period Percentage of Matched GDPR and non-GDPR Websites that Update** .................................................. 33

5.7 **Average Period-by-Period Update Significance for GDPR and non-GDPR Websites** .................................................. 35

6.1 **Accessibility Measures: Average Length and Readability of GDPR and non-GDPR Privacy Policies** ............................................... 39

6.2 **DPO Contact: Percentage of GDPR Privacy Policies that Include Contact Details** .................................................. 42

6.3 **Data Subject Rights: Percentage of GDPR Privacy Policies that Mention Data Subject Rights** .................................................. 43

6.4 **Legal Basis of Processing: Percentage of GDPR Privacy Policies that Mention A Legal Basis of Processing** ............................................. 44
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td><strong>Summary Statistics</strong>: GDPR and non-GDPR Covariates ($X$)</td>
<td>26</td>
</tr>
<tr>
<td>5.1</td>
<td><strong>Wilcoxon Signed-Rank Test</strong>: Update Probability Comparison between Matched GDPR and non-GDPR Populations</td>
<td>34</td>
</tr>
<tr>
<td>5.2</td>
<td><strong>Wilcoxon Signed-Rank Test</strong>: Update Significance Comparison between Matched GDPR and non-GDPR Populations</td>
<td>36</td>
</tr>
<tr>
<td>6.1</td>
<td><strong>Wilcoxon Signed-Rank Test</strong> ($Y_{length}$): Average Length Comparison of GDPR and non-GDPR Privacy Policies</td>
<td>40</td>
</tr>
<tr>
<td>6.2</td>
<td><strong>Wilcoxon Signed-Rank Test</strong> ($Y_{flesh}$): Average Readability Comparison of GDPR and non-GDPR Privacy Policies</td>
<td>40</td>
</tr>
<tr>
<td>B.1</td>
<td>“Website” Database Sample</td>
<td>55</td>
</tr>
<tr>
<td>B.2</td>
<td>“Privacy Policy” Database Sample</td>
<td>56</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The Internet has historically been regarded as a “lawless space, both un gover ned in practice and un gov ernable by design” [1]. Market-based forces based on principles of “notice and choice”\(^1\) have long been the dominant regulatory framework of Internet: websites “notify” consumers about what personal data they are collecting in privacy policies, and consumers then “choose” whether to sell ownership of their data on the free market [3].

Adopted in April 2016 and passed in March 2018, the General Data Protection Regulation (GDPR) offers a first glimpse of what a strong Internet privacy law may look like [1]. Broadly speaking, the GDPR has been the farthest-reaching privacy regulation in history\(^2\) and has greatly strengthened privacy policy guidelines. GDPR Article 12 requires that all websites have a “concise, transparent, intelligible and easily accessible privacy policy, written in clear and plain language.” Various additional GDPR Articles require that privacy policies mention a Data Protection Officer contact address (GDPR Article 13b), the legal basis of all personal

\(^1\)Market-based digital regulations have long been controversial, and legal scholars have extensively debated the merits of “notice and choice” as discussed in Appendix A.1 [2].

\(^2\)The GDPR applies both to all websites located in the European Union and to all websites outside of the European Union that collect personal data from European citizens. Previously, a Safe Harbor Agreement between the European Court of Justice and US Department of Commerce found that US “notice and choice” offered sufficient privacy protections and permitted US companies to collect personal data on European citizens without adopting the European Data Protection Directive (95/46/EC) [4]. In 2016, the European Court of Justice repealed the Safe Harbor Agreement and enacted a more stringent EU-US Privacy Shield Framework that requires all United States websites to abide by the GDPR to process European data [5].
data processing (GDPR Article 13a), and various data subject rights (GDPR Chapter III)\(^6\). A more comprehensive overview of all GDPR regulations can be found in Appendix A.2.

GDPR regulation has inspired many successive privacy developments and has tremendous potential to strengthen digital privacy reforms\(^4\). However, the GDPR’s causal impact has never been formally analyzed. Many countries including Brazil, South Korea, Argentina, New Zealand, India, and South Africa are blindly modeling privacy laws off of the GDPR without a clear understanding of GDPR successes and limitations. To my knowledge, this thesis is the first to quantify the causal impact of GDPR regulation on privacy policies. All previously published privacy policy databases are ill-suited for causal inference: all aggregate GDPR and non-GDPR privacy policies and are limited to short observation periods. No prior research has been able to assess causality as a result. To overcome data limitations, I curate both a “website” dataset of 31,843 distinct websites classified as GDPR and non-GDPR with 88.5% accuracy and a large-scale, longitudinal “privacy policy” dataset of 317,396 historic privacy policies from 2004 to 2019 corresponding to each of the 31,843 websites. I use both datasets to compare GDPR and non-GDPR websites’ privacy policies before and after GDPR adoption to establish causality.

This analysis is divided into two parts. First, I assess whether the GDPR has changed privacy policy content using a matched difference-in-difference causal estimator. Results indicate that GDPR regulation has caused GDPR privacy policies to change 12.45% more than a non-GDPR control. I explore nuances of how the GDPR caused changes in privacy policies and find that the GDPR changed privacy policies by causing more websites to update more frequently. GDPR regulation has caused 25% more websites to update privacy policies than otherwise would have, ceteris paribus, and has caused GDPR websites to update their privacy policies 25.43% more frequently.

Second, I assess how the GDPR affects privacy policy readability and compliance. On

\(^3\)Article 15 guarantees the “right of access by a data subject.” Article 16 guarantees the right to rectification. Article 17 guarantees the right to erasure ('right to be forgotten'). Article 18 guarantees the right to restrict processing. Article 20 guarantees the right to data portability.
one hand, my findings indicate that the GDPR has negatively impacted privacy policy readability: GDPR privacy policies have become 67.85% longer and 0.36 grade-levels less readability after GDPR adoption. On the other hand, results also indicate that the GDPR has positively impacted privacy policy compliance with various GDPR Articles. 15% more GDPR privacy policies mention a Data Protection Officer’s contact information and 23% more mention a legal basis for personal data processing after GDPR adoption. An inherent tension appears to exist between accessibility and compliance in privacy policies. GDPR privacy policies have become more detailed to comply with various GDPR Articles but less readable as a result. Additional research is needed on how to balance this seemingly intractable trade-off between accessibility and transparency.

Considered in tandem, the results from both analyses underscore the pressing need for further research on GDPR success and limitations. The GDPR has changed privacy policies. However, it is less clear that the GDPR has positively changed privacy policies. My results indicate that if the GDPR is to successfully serve as a model for future regulatory developments, policymakers need to have a clearer understanding of the GDPR’s nuanced effects on privacy policies.

The remainder of this thesis proceeds as follows. Chapter 2 summarizes related research on privacy policies, and Chapter 3 describes the statistical causal methods used throughout all analyses. Chapter 4 details the creation of a longitudinal privacy policies database. Chapter 5 explores quantitative questions of whether GDPR regulation has significantly changed privacy policy content. Chapter 6 explores more qualitative questions of how GDPR regulation has changed policy content. Chapter 7 offers concluding remarks.
Chapter 2

Related Work

This chapter synthesizes prior research and explains how this thesis connects to a large literature on privacy policies.

2.1 Cross-Sectional

2.1.1 Privacy Policy Adoption

Prior cross-sectional research has analyzed the adoption of privacy policies at various points in time. No law requires all websites to post a privacy policy in the United States\(^1\). A Federal Trade Commission (FTC) Report (1998) analyzed a set of 674 United States websites and found that 92% collected personal user data but only 14% posted a privacy policy \(^8\). Liu (2002) analyzed a different set of 600 Top 500 stock-exchanged websites and found that the number of websites that posted a privacy policy increased to 30% by 2002 \(^4\). Most recently, Story (2018) analyzed a third set of 1,035,853 Android apps and found that 51.8% had a posted privacy policy by May 2018 \(^9\).

This thesis importantly builds upon previous “privacy policy adoption” research through

\(^1\)No general federal law in the United States requires all websites to have a privacy policy. However, taken together, the Children’s Online Privacy Protection Act and the California Online Privacy Protection Act of 2003 de facto require privacy policies for most websites \(^7\).
a longitudinal rather than cross-sectional analysis. All previous research has been conducted on different subsets of websites at a specific moment in time. A comparison of different datasets fails to control for hidden bias and may lead to invalid conclusions. (Liu’s 600 Top 500 stock-exchanged websites and Story’s 1,035,853 Android apps, for example, are likely fundamentally different comparison groups.) This thesis analyzes how the GDPR has changed privacy policy adoption rates on the same subset of websites over time in Chapter 5.

2.1.2 Privacy Policy Deficiencies

Another cross-section research area analyzes privacy policy deficiencies due to poor accessibility and poor transparency [10, 11]. Prior research on accessibility shows that privacy policies are unreasonably long and difficult to understand. McDonald (2004) found that it takes users, on average, 11 minutes to read a website’s privacy policy and 84 minutes to read a website and all associated third-party privacy policies [12]. Otherwise stated, it would take users 244 hours annually to read the privacy policy of every website they visit. Ermakova (2015) further found that the average privacy policy requires a college-reading level to understand [13].

Prior research on transparency additionally shows that privacy policies are vague and often misleading. Reidenberg (2016) analyzed a sample of 15 privacy policies and found that 70.6% contain vague terminology (e.g., “may”, “as necessary”, “at our discretion”, etc.) [14]. Okoyomon analyzed a set of 9,424 app privacy policies and found numerous deceptive ambiguities. 28.4% of apps that are un-encrypted-in-transit misleadingly claimed to “take measures to secure data transfer.” 10.5% of apps that share data with third-party services fail to mention that in their privacy policies. 87.8% of apps that share data with third-party services fail to name the third-parties they are sharing with and instead hide under

---

2These calculations assume a standard reading rate of 250 words per minute and an average user visits 94 websites per day.
ambiguous statements like, “We are not responsible or liable for third party data practices” [15]. Poor accessibility and poor transparency both hamper consumer comprehension of privacy policies and cause many consumers to stop reading privacy policies altogether [16]. Obar (2018) found that 74% of consumers fail to even click on a privacy policy before joining a site [17].

Again, this thesis importantly builds upon previous “privacy policy deficiency” research by using a longitudinal rather than cross-sectional approach. All previous work only analyzes privacy policy accessibility and transparency at specific “snapshots” in time. A central aim of the GDPR has been to improve privacy policy notice. Chapter 6 analyzes how successfully GDPR has changed both accessibility and transparency over time. Results from Chapter 6 identify an apparent tension between privacy policy deficiencies and suggest that the GDPR may be unable to improve both accessibility and compliance at the same time.

2.2 Longitudinal

Concurrent to this thesis, a few prior studies used longitudinal privacy policy analysis to study the effects of the GDPR. Degeling (2019) performed the first large-scale longitudinal analysis of GDPR impact on 6,579 privacy policies from 24 different European countries and found that the percentage of websites with privacy policies increased 4.9% after the GDPR [18]. Linden (2019) conducted a similar analysis of 6,278 European privacy policies and found that more privacy policies include mention of advanced tracking technologies after the GDPR [19]. Lastly, Amos (2020) performed a significantly larger-scale analysis of 1,071,480 privacy policies and found that privacy policies have become longer and less comprehensible since GDPR adoption [20].

This thesis builds upon prior work in two important ways. First, all previous studies compare pre-GDPR privacy policies from immediately before GDPR adoption (March 2018) to post-GDPR privacy policies from immediately after GDPR adoption (September 2018).
This thesis analyzes privacy policies over a significantly longer observation period (2004 to 2019) to characterize longer-term GDPR effects. Second, all previous studies only consider privacy policies under GDPR regulation and fail to compare GDPR privacy policies to a non-GDPR control. All suggest different correlations between GDPR adoption and privacy policy changes, but none are able to prove causality. This thesis is the first to analyze the GDPR’s causal impact on privacy policies against a non-GDPR control.
Chapter 3

Methods

This chapter describes causal statistical methods used in Chapters 5 and 6. Specifically, this thesis uses difference-in-difference and Kaplan Meier survival estimators to quantify the causal impact of the GDPR.

3.1 Basic Notation

All causal analysis in this thesis follows a Neyman-Rubin potential outcomes framework. Consider a website $i$. Let $Y_i^{\text{pre}}$ be the website’s outcome before GDPR adoption and $Y_i^{\text{post}}$ be the website’s outcome after GDPR adoption.\(^1\) The Neyman-Rubin causal framework assumes that the website $i$ will experience one of two potential outcomes: $Y_i(1)$ represents an outcome that would be observed if the website were under GDPR jurisdiction and $Y_i(0)$ represents an outcome that would be observed if the website were not under GDPR jurisdiction. Let $T$ represent a binary GDPR indicator variable ($T_i = 1$ for GDPR websites, $T_i = 0$ for non-GDPR websites).

\(^1\)This chapter draws heavily from Imbens and Rubin’s *Causal Inference for Statistics, Social, and Biomedical Sciences* and Hernan and Robins’s *Causal Inference: What If* \cite{imbens2015causal, hernan2020causal}.

\(^2\)This thesis quantifies GDPR causal impact before and after the GDPR adoption period rather than the GDPR enforcement period. The GDPR was adopted on April 14, 2016, after which websites began adapting their privacy policies to published GDPR standards. The GDPR did not become enforceable until May 25th, 2018, after which non-compliant websites could be fined.
The Neyman-Rubin causal framework implies that a website $i$ has four potential outcomes:

$$Y_{i}^{\text{pre}}(0), \ Y_{i}^{\text{pre}}(1), \ Y_{i}^{\text{post}}(0), \ Y_{i}^{\text{post}}(1).$$

A website’s observed outcome before GDPR adoption ($Y_{i}^{\text{pre}}$) equals both potential outcomes assuming that future treatment does not affect past outcomes:

$$Y_{i}^{\text{pre}} = Y_{i}^{\text{pre}}(0) = Y_{i}^{\text{pre}}(1). \quad (3.1)$$

A website’s observed outcome after GDPR adoption ($Y_{i}^{\text{post}}$) equals

$$Y_{i}^{\text{post}} = T \cdot Y_{i}^{\text{post}}(1) + (1 - T) \cdot Y_{i}^{\text{post}}(0). \quad (3.2)$$

Let $X_{i}$ be a website’s covariate vector that does not change over time. Let the propensity score $\pi(X)$ represent the conditional probability of GDPR treatment assignment given covariate vector $X_{i}$:

$$\pi(X_{i}) = Pr(T_{i} = 1 | X_{i}). \quad (3.3)$$

The two most common estimands in Neyman-Rubin causal analysis are average treatment effect (ATE) and average treatment effect on treated (ATT). ATE measures the average treatment effect on a population as

$$\tau_{ATE} = E(Y_{i}^{\text{post}}(1) - Y_{i}^{\text{post}}(0)) \quad (3.4)$$

and ATT measures the average treatment effect on only the treated group as

$$\tau_{ATT} = E(Y_{i}^{\text{post}}(1) - Y_{i}^{\text{post}}(0) | T_{i} = 1). \quad (3.5)$$

---

I only consider pre-GDPR covariates to avoid post-treatment covariate bias. Thus, $X_{i} = X_{i}^{\text{pre}}$. 

9
Otherwise stated, $ATE$ answers the question, “What would difference in outcomes ($Y$) would we expect to observe if GDPR regulation applied to all websites compared to no websites?” $ATT$ addresses the question, “What difference in outcomes ($Y$) would we expect to observe if all GDPR websites were instead non-GDPR websites?” This thesis quantifies the GDPR $ATT$ to understand the actual impact that GDPR regulation has had on websites rather than the hypothetical impact that GDPR regulation would have had if applied to all websites.

### 3.2 Assumptions

All causal analysis in this thesis depends upon three assumptions: Stable Unit Treatment Value, Probabilistic Treatment Assignment, and Unconfounded Treatment Assignment.

**Stable Unit Treatment Value**  Stable Unit Treatment Value (SUTVA) assumes both no hidden treatment variation and no interference between units. No hidden treatment variation assumes that GDPR regulation is uniformly applied to all GDPR websites (e.g., GDPR regulation should not vary geographically across different local Data Protection Authorities). No interference assumes that a website’s potential outcomes do not depend on whether another website is under GDPR jurisdiction (e.g., the number of websites under GDPR jurisdiction should not affect outcomes).

The next two assumptions concern a key component of causal analysis: the treatment assignment mechanism. A treatment assignment mechanism represents the process by which units are assigned to treatment or control. Formally, the assignment mechanism $Pr(T|X,Y^0,Y^1)$ assigns probabilities to all $2^N$ possible treatment assignments of $N$ units.
The individual treatment assignment probability for website $i$ may then be calculated as

$$\pi(X_i) = \sum_{T:T_i=1} Pr(T | X, Y, Y_0, Y_1). \quad (3.6)$$

Randomized trials are the gold standard of causal inference and have a known and carefully controlled treatment assignment mechanism. Observational studies pose a more difficult challenge: their treatment assignment mechanisms are not necessarily known and uncontrolled by researchers. Thus, I assume that the GDPR treatment assignment mechanism is both probabilistic and unconfounded throughout this observational analysis in order to draw causal conclusions.

**Probabilistic Treatment Assignment**  A treatment assignment is probabilistic if all websites have a positive probability of being assigned all $T \in \{0, 1\}$:

$$0 < \pi(X_i) < 1. \quad (3.7)$$

**Unconfounded Treatment Assignment**  A treatment assignment is unconfounded if treatment is independent of potential outcomes conditional on covariates $X_i$:

$$P(T_i|X_i, Y_i(0), Y_i(1)) = P(T_i|X_i) = \pi(X_i) \quad (3.8)$$

or, in the Dawid (1979) conditional independence notation,

$$(Y_i^0, Y_i^1) \perp T_i | X_i. \quad (3.9)$$
Unconfoundedness is an incredibly important assumption in observational studies and allows the causal effect \( \tau \), conditional on covariates \( X_i \), to be rewritten as

\[
\tau_{ATE} = E(Y_i(1) - Y_i(0) \mid X_i)
\]

\[
= E(Y_i \mid T_i = 1, X_i) - E(Y_i \mid T_i = 0, X_i).
\]

This result is later used by the difference-in-difference estimator.

Unconfoundedness is also an incredibly difficult assumption to verify in observational studies. Researchers can never definitively prove both \( Y_i(0) \) and \( Y_i(1) \) are both independent of treatment assignment since at most one potential outcome can be observed for website \( i \). A statistical technique called Rosenbaum sensitivity analysis assesses the robustness of outcomes to potential unobserved confounders and quantifies how strong an unmeasured covariate would have to be to alter causal conclusions. Formally, Rosenbaum sensitivity analysis considers two units, \( j \) and \( k \), with the same observed covariate values \( (x_j = x_k) \) but with different propensity scores \( (\pi_j \neq \pi_k) \) due potentially to an unobserved covariate. The odds ratio of unit \( j \) receiving treatment compared to unit \( k \) is bound by \( \Gamma \), for \( \Gamma \geq 1 \), such that:

\[
\frac{1}{\Gamma} \leq \frac{\pi_j}{1-\pi_j} \leq \Gamma.
\]

Simply stated, \( \Gamma \) represents the minimum strength of association that an unmeasured confounder would need to have to fully explain away a causal association, conditional on the measured covariates. If \( \Gamma = 1 \), then all units with similar covariates \( (x_j = x_k) \) would have the same probability of treatment \( (\pi_j = \pi_k) \), and no hidden bias would exist. If \( \Gamma = 2 \), unit \( j \) would be twice as likely as unit \( k \) to receive treatment due to potential unobserved covariate differences, and a hidden bias would exist. Rosenbaum sensitivity analysis systematically increases \( \Gamma \) to calculate how much hidden bias can be present – how large \( \Gamma \) can be – before causal conclusions. A small \( \Gamma \) implies few unmeasured confounders could explain away an estimated causal effect. A large \( \Gamma \) implies many unmeasured confounders would be needed to
explain away an estimated causal effect. No prior research offers a specific threshold cut-off for \( \Gamma \), but unconfoundedness is a more plausible assumption for larger \( \Gamma \). Chapter 5 uses a Rosenbaum sensitivity analysis to assess unconfoundedness between GDPR treatment and outcomes.

### 3.3 Difference-in-Difference Estimator

Throughout Chapters 5 and 6, I first use matched difference-in-difference estimators to estimate GDPR \( ATT \) on various outcomes \( Y \). Difference-in-difference estimators adjust both for pre-GDPR and post-GDPR time trends and unobserved differences between GDPR and non-GDPR websites. Formally, a difference-in-difference estimator \( (\tau_{ATT}^{DOD}) \) is defined as:

\[
\tau_{ATT}^{DOD} = (E(Y_{i}^{post}|T_i = 1) - E(Y_{i}^{pre}|T_i = 1)) - (E(Y_{i}^{post}|T_i = 0) - E(Y_{i}^{pre}|T_i = 0)). \tag{3.13}
\]

This thesis matches GDPR and non-GDPR websites and then estimates \( \tau_{ATT}^{DOD} \) by replacing all terms in Equation 3.13 with the conditional sample means. Matching is a necessary non-parametric statistical method that pairs treatment and control units with similar pre-GDPR outcomes \( (Y_{i}^{pre}) \) and controls for confounding bias. Intuitively, if GDPR and non-GDPR websites systematically differ, observing \( Y_{i}(0) \) only for non-GDPR websites does not inform our understanding of \( Y_{i}(0) \) for GDPR websites. Matching validates a “parallel trends” assumption that assumes GDPR and non-GDPR websites’ mean outcomes would have followed parallel trends without GDPR regulation \[23\]:

\[
E(Y_{i}^{post}(0) - Y_{i}^{pre}(0)|T_i = 0, X_i) = E(Y_{i}^{post}(0) - Y_{i}^{pre}(0)|T_i = 1, X_i). \tag{3.14}
\]

Chapter 4 matches GDPR and non-GDPR websites in a random order, without-replacement,
using a 1NN matching algorithm\footnote{4}

\[ m(i) = \arg\min_{j: D_j \neq D_i} K(X_j, X_i). \] (3.15)

I define \( K(x, x') \) to be the Mahalanobis distance function\footnote{5}

\[ K(x, x') = \exp(- (x - x)^T W (x - x')) = (x - x')^T \left( \frac{N_c \hat{\Sigma}_c + N_t \hat{\Sigma}_t}{N_c + N_t} \right)^{-1} (x - x')^T, \] (3.17)

where \( \hat{\Sigma}_c \) and \( \hat{\Sigma}_t \) represent the within-sample covariance matrices of GDPR and non-GDPR websites, respectively. A within-sample covariance matrix (\( \hat{\Sigma}() \)) is calculated as

\[ \hat{\Sigma}() = \frac{1}{N() - 1} \sum_{i: T_i = ()} (X_i - \bar{X}())(X_i - \bar{X}())^T. \] (3.18)

### 3.4 Survival Analysis

Lastly, in Chapter\footnote{3} I use survival analysis to estimate GDPR causal impact on time until a website updates its privacy policy. Observational studies pose an intrinsic problem of right-censoring. Websites may not update their privacy policy in the observation period and, thus, may have an unknown time until update \( D \). Previous difference-in-difference estimators are ill-suited to handle censored data. Survival analysis importantly accounts for censoring and quantifies GDPR effect on time until an outcome \cite{21}.

Let \( D \) be a non-negative, continuous random variable that represents time until a website updates its privacy policy.
updates its privacy policy. Let $C$ denote a dichotomous censorship variable ($C = 0$ for censored outcome, $C = 1$ for observed outcomes). The \textit{survival function} $S(t)$ represents the probability that a website has not updated its privacy policy by time $t$:

$$S(t) = 1 - F(t). \quad (3.19)$$

The \textit{cumulative hazard function} $H(t)$ represents the probability that an update has occurred by time $t$:

$$H(t) = F(t) = \int_0^t \lambda(x) \, dx. \quad (3.20)$$

Lastly, the \textit{hazard function} $\lambda(t)$ represents the instantaneous probability that a website updates its privacy policy at a given time $t$:

$$\lambda(t) = Pr(D = t | D \geq t). \quad (3.21)$$

To quantify GDPR $AT\tau \quad (\tau_\text{ATT})$ on time until a privacy policy update, I first use a non-parametric Kaplan–Meier estimator\footnote{Note that a Nelson-Aalen estimator offers a closely related non-parametric estimate of GDPR and non-GDPR of \textit{cumulative hazard} $H(t)$:}

\begin{equation}
\hat{S}(t) = \prod_{i: t_i \leq t} \frac{n_i - d_i}{n_i}, \quad (3.23)
\end{equation}

where $d_i$ represents the total number of websites that update their privacy policy at time $t$, and $n_i$ represents the total number of websites that have not updated their privacy policy by time $t$. I then directly compare matched GDPR and non-GDPR websites’ survival curves $S(t)$ to understand GDPR impact on time until a first update. Recall matching techniques from Section\textsuperscript{3.3} control for confounding bias between GDPR and non-GDPR websites and

\begin{equation}
\hat{H}_{\text{NA}}(t) = \sum_{i: t_i \leq t} \frac{d_i}{n_i}, \quad (3.22)
\end{equation}

Both Kaplan-Meier and Nelson-Aalen curves are included in Chapter\textsuperscript{5}.\footnote{Both Kaplan-Meier and Nelson-Aalen curves are included in Chapter\textsuperscript{5}}
validate Kaplan-Meier analysis.

Lastly, I use cure to estimate the portion of GDPR and non-GDPR websites expected to never update their privacy policy. I define a ‘cured’ website as a website that will never update its privacy policy or that will update its privacy policy so far in the future that its time until event is essentially infinity \( d_i = \infty \). To quantify GDPR and non-GDPR cure population, I define a new cure cumulative hazard \( H(t) \) as

\[
H_{\text{cure}}(t; c, \theta) = cH(t; \theta)
\]  

where \( c \) represents an unknown horizontal cure asymptote and \( H(t) \) represents a previously-defined parametric cumulative hazard of uncured units. I then quantify the proportion of GDPR and non-GDPR websites expected to never update their privacy policies \( c \) by applying an Upper Asymptote Fitter to \( H_{\text{cure}}(t; c, \theta) \) using the Lifelines Python library [25].
Chapter 4

Database

This chapter describes the methodologies and challenges of curating a large-scale privacy policy database. The final database consists of two datasets: a “website” dataset and a “privacy policy” dataset. The “website” dataset contains 31,843 GDPR and non-GDPR websites classified with 88.5% accuracy. The “privacy policy” dataset contains 317,396 historical privacy policies from 2004 to 2019 for all 31,843 websites. The “privacy policy” database is discretized into semi-annual periods from 2004 to 2019. Period A(·) represents the first half of a year (January-June) and Period B(·) represents the second half of a year (July to December). A sample of both datasets is provided in Appendix B.2. The two datasets allow me to compare GDPR and non-GDPR websites’ privacy policies both before and after GDPR adoption to assess causality later in Chapters 5 and 6.

4.1 Website Dataset

The “website” dataset consists of a set of 31,843 GDPR and non-GDPR websites. Note that all “website” dataset rows correspond to a unique website. The “website” dataset also defines four covariate columns First Year ($X_0$), Duration ($X_1$), Category ($X_2$), and Rank ($X_3$) defined below. These covariate columns are later used to match GDPR and non-GDPR websites in Section 4.4.
• **First Year** ($X_0$). First Year represents the date of a website’s earliest archived homepage.

• **Duration** ($X_1$). Duration represents the number of years a website has had an achieved Wayback Machine homepage. Both *First Year* and *Duration* are calculated through the Wayback Machine API as `oldest().archive_url` and `newest().archive_url - First Year`, respectively.

• **Category** ($X_2$). Category represents a website’s business category and is calculated using the Webshrinker Category API[1].

• **Rank** ($X_3$). Rank represents a website’s average annual Alexa rank from 2004 to 2019 and is calculated using the Alexa TopSites API [26].

To create the “website dataset”, I first pull the daily archived Alexa Global Top 1M lists for each observation midpoints from 2004 to 2019 and selected over 75,550 websites in the top 50K of a recent Alexa lists [26] [27]. I consider 50K a large enough threshold to cover both GDPR and non-GDPR websites. I then use the Wayback Machine wayback-cdx-server[2] to retrieve archived homepages for all of the 75,551 websites [29]. A sample CDX Server GET Request is shown in Appendix B.1. I restrict GET requests to midpoint dates of observation periods to avoid potential computational challenges[3]. I filter achieved homepages by English language (`language = ’en’`) using the Polyglot Python library [31]. (Analyzing non-English privacy policies would require language proficiency beyond the scope of this thesis.) I also filter archived homepages by `status_code = 200` and `mimetype = text/html`.

---

[1] I map a website belonging to multiple categories to its first (most prominent) category. The most frequent categories are `{"business", "education", "entertainment", "uncategorized", "shopping", "adult", "information tech", "games"}`.

[2] The Wayback Machine is a digital archive that records billions of historic website homepages. The wayback-cdx-server is a standalone Wayback Machine HTTP servlet that returns an achieved homepage capture as a `<urlkey, timestamp, original url, mimetype, status code, digest, length>` vector [28].

[3] An unfiltered CDX Server GET request may return billions of rows depending on how frequently the Wayback Machine achieved that domain. The Wayback Machine can archive up to 7,200 captures per day for popular websites [30].
The resulting dataset contained 41,616 websites that altogether spanned 407,837 archived homepages captures.

Lastly, I remove all parked and cross-origin homepage redirect (COHR) websites from the dataset as recommended by Amos [20]. A parking website is a website that is registered but redirects to a dummy URL. A cross-origin homepage redirect (COHR) website is a website that redirects to a different domain.[4] I manually removed all parked websites using an ICANN-Accredited Registrars list. I manually removed all COHR websites by removing rows with different base URLs but equal covariate columns ($\{X_{i,0}, X_{i,1}, X_{i,2}, X_{i,3}\}$).[5]

4.2 Privacy Policy Dataset

The “privacy policy” dataset consists of archived privacy policies from 2004 to 2019 for all websites in the “websites” dataset. Note that the “privacy policy” dataset is longitudinal: the same website’s privacy policy will be recorded multiple times for every semi-annual observation period between 2004 and 2019. To create the “privacy policy” dataset, I match the previously-collected achieved homepages to their corresponding privacy policies in the Princeton-Leuven Longitudinal Privacy Policies Corpus on date and URL fields [32]. Appendix B.3 describes the Princeton-Leuven Longitudinal Privacy Policies Corpus in greater detail. The final “privacy policy” dataset contains 31,843 GDPR privacy policies from 2004 to 2019.

The “privacy policy” dataset also includes four additional privacy policy text-feature columns $fuz$, $update$, $length$, and $readability$ defined below. Chapter 5 will analyze GDPR causal impact on $fuz$ and $update$, and Chapter 6 will analyze GDPR causal impact on both $length$ and $readability$.

A parked domain often occurs when a large company registers a domain name to reserve naming rights and leaves the domain out of service. A COHR often occurs when a large company redirects perceived misspelled domains to the correct location (e.g., “exmaple.com” redirects to “example.com”). A COHR may also occur in malicious phishing or malware attacks.

I use base URLs rather than complete URLs so that “example.net” redirecting to “example.com” would not be removed as a COHR.
• **Fuz.** Fuz represents the percentage a privacy policy’s text has not changed between the current period \( p \) and the immediately prior period \( p^* \). Both Fuz and Update calculations are described below.

• **Update.** An indicator variable Update represents whether a privacy policy has *significantly* changed in a given period such that the current privacy policy and previous version differ by more than 5% \( (Fuz < 0.95) \).

• **Length.** Length represents a privacy policy’s character count and is calculated using the Python Standard library.

• **Readability.** Readability represents a privacy policy’s readability and is calculated using the Flesch–Kincaid Grade Level formula:

\[
Y_{flesh} = 0.39 \left( \frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left( \frac{\text{total syllables}}{\text{total words}} \right) - 15.59. \tag{4.1}
\]

Note that a higher Flesch–Kincaid Grade Level \( (Y_{FKGL}) \) signifies worse readability. Originally developed in 1976 for the U.S. Navy, Flesch-Kinkaid Grade Level has become the most widely used readability formula in the United States. However, minimal explanation exists on how Flesch-Kinkaid Grade Level weighting factors 0.39 and 11.8 were determined.

Before calculating all four text-feature columns, I preprocess privacy policies. I remove unnecessary HTML boilerplate using the HTML Boilerpipe extraction library \(^{33}\) and extract Markdown-formatted text using the Beautiful Soup Python library \(^{34}\). I remove non-sentence headers, tables, and lists using the Mistune Python library. I then convert all text to lowercase, removed all punctuation using the Python string library, and manually removed a limited list of stop words\(^6\).

---

\(^6\)I hand-defined a limited list of stop words (e.g., ‘the’, ‘is’, ‘as’, ‘a’, ‘are’, ‘in’, ‘this’, ‘that’) rather than relying on larger NLTK Python libraries to avoid biasing future analysis. For example, removing the stop word ‘not” from the sentence ”We do not sell any of our location data” may significantly alter meaning.
I then calculate \( fuz \) and update metrics using the FuzzyWuzzy Python library. FuzzyWuzzy is a modified measure of Levenshtein distance. Levenshtein distance is defined as the minimum number of single-character edits (i.e., insertions, deletions, or substitutions) required to convert string \( a \) into string \( b \) and is formally defined as

\[
lev(a, b) = \begin{cases} 
|a| & \text{if } |b| = 0 \\
|b| & \text{if } |a| = 0 \\
lev(tail(a), tail(b)) & \text{if } a[0] = b[0] \\
1 + min \begin{cases} 
lev(tail(a), b) \\
lev(a, tail(b)) \\
lev(tail(a), tail(b))
\end{cases} & \text{otherwise.}
\end{cases}
\] (4.2)

The FuzzyWuzzy similarity ratio modifies Levenshtein distance by normalizing it to a bounded ratio between \([0, 1]\):

\[
fuz(a, b) = 1.0 - \frac{lev(a, b)}{\max(|a|, |b|)}
\] (4.3)

where \(|a|\) and \(|b|\) are the lengths of \( a \) and \( b \) respectively. (Appendix B.4 walks through an example Fuzzy Similarity calculation.) A ratio of 1 indicates that two strings are identical, and a ratio of 0 indicates that two strings share no similarities. Notice that the Levenshtein algorithm is computationally taxing (i.e., \( O(|a||b|) \) time complexity, \( O(|a|) \) space complexity). The FuzzyWuzzy Python library optimizes Levenshtein distance to larger strings by tokenizing the strings into sentences and then computing the average FuzzyWuzzy similarity ratio between differing sentences rather than both strings. Alternative, less computationally-taxing metrics I considered were Jaro–Winkler and cosine similarity, but both fail to account for word order and perform worse on the dataset [35, 36].

Specifically, I calculate \( fuz(p, p^*) \) as the period-to-period FuzzyWuzzy similarity between
a privacy policy at period $p$ and its previous version at $p^*$. I define $I_{update}$ as an indicator of whether a website has *significantly* changed its privacy policy in a given period $p$:

$$I_{update}(t) = \begin{cases} 
0 & \text{if } fuz(p, p - 1) \geq 95 \\
1 & \text{if } fuz(p, p - 1) < 95.
\end{cases}$$ (4.4)

### 4.3 GDPR Classification

The “websites” database also includes a column $GDPR (X_4)$ that represents whether a website is under GDPR jurisdiction. I manually classify all websites in the “website database” as GDPR and non-GDPR using a combined country code top-level domain (ccTLD) and privacy policy “tip-off” term approach. Recall that the GDPR applies both to all websites located in Europe and to all websites that collect or process personal information from European residents [6]. To begin, I classify all websites located in Europe that contain a European country code top-level domain (ccTLD) (i.e., .eu, .fr, .ch, .dk, .it, .hu, etc.) as GDPR websites. ccTLD classification alone has high precision but limited scope [37]. Many European websites may not have a European ccTLD (only 40% of websites use ccTLDs), and the GDPR applies to many websites located outside of the European Union that do not have a European ccTLD. Therefore, to augment ccTLD classification, I search privacy policies for the GDPR “tip-off” terms using the spaCy Python library, and then I classify any website that has a privacy policy with a “tip-off” term as GDPR. Ultimately, the “website” dataset has 12,110 GDPR websites and 19,733 non-GDPR websites. The GDPR set is markedly smaller than the non-GDPR set in part because the Wayback Machine crawler excluded a significant portion of European-language GDPR websites as mentioned above.

Next, I assess the accuracy of this classification on a random sample of 100 GDPR and 100 non-GDPR websites. Non-GDPR websites are required to block all European traffic.

---

7 Note that $E(I_{update}) = P(update)$ by Linearity and the Fundamental Bridge.

8 Linden (2019) has identified various “tip-off” terms that signal GDPR compliance (i.e., “GDPR”, “Safe Harbor”, “Privacy Shield”, “UK US agreement”, “Europe”, “European”) [19].
Assuming perfect compliance, all websites that can be accessed from a European IP address should be GDPR. I temporarily connect to each website from a Spanish server via a VPN to check whether the website is geo-blocked and establish ground truth (i.e., whether a website really is GDPR or non-GDPR). A confusion matrix of my classification’s performance on the sample of 200 privacy policies is shown in Figure 4.1. My classification appears to maximize precision over recall and be overlooking a marked number of GDPR websites. All websites classified as GDPR are GDPR and successfully allow a German IP address to access them (Equation 4.5) but 23% of non-GDPR classified website are likely GDPR (Equation 4.6). This is likely because my classification begins with the blanket assumption that all websites are non-GDPR and then hand-selects GDPR websites, presumably missing many. GDPR classification has high enough accuracy (88.5%) (Equation 4.7) to support causal conclusions. Chapters 5 and 6 use this classification to compare the causal impact of GDPR regulation across GDPR and non-GDPR control privacy policies. However, additional research may significantly refine classification techniques as discussed in Section 4.5.

\[
\text{precision} = \frac{TP}{TP + FP} = 1.0 \quad \text{(4.5)} \\
\text{recall} = \frac{TP}{TP + FN} = .813 \quad \text{(4.6)} \\
\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = .885 \quad \text{(4.7)}
\]
**Figure 4.1: GDPR Classifier Confusion Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Predicted(^{(1)})</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GDPR</td>
<td>Non-GDPR</td>
</tr>
<tr>
<td>Actual(^{(1)})</td>
<td>100 (TP)</td>
<td>23 (FN)(^{(2)})</td>
</tr>
<tr>
<td>Non-GDPR</td>
<td>0 (FP)(^{(3)})</td>
<td>77 (TN)</td>
</tr>
</tbody>
</table>

Notes:

\(^{(1)}\) *Predicted* is determined through manual classification described above. By design, 100 websites in the sample are classified as GDPR, and 100 are classified as non-GDPR. *Actual* is determined through geo-blocking verification described above. 123 websites in the sample actually are GDPR, and 77 actually are non-GDPR.

\(^{(2)}\) *False Positive (FP)* represents websites that are manually classified as GDPR but do *not* allow access to a Spanish IP address (implying they are not GDPR sites). *False Negative (FN)* represents websites that are manually classified as non-GDPR but *do* allow access to a Spanish IP address (implying they are GDPR sites).
4.4 Matching

I lastly create a matched subset of GDPR and non-GDPR websites from the “websites” database used throughout Chapter 5 and Chapter 6. To begin, Table 4.1 shows that GDPR and non-GDPR websites exhibit significant underlying covariate differences in the “websites” dataset: GDPR websites are, on average, 1.03 years younger and have a 4,028 position-higher Alexa rank than non-GDPR websites. All $t$-tests yield $p$-values greater than 0.05. The results of Table 4.1 indicate that GDPR and non-GDPR covariate differences may be so large that simple regression adjustments may not adequately control for them. Matching is needed to balance GDPR and non-GDPR covariate differences.

I match GDPR and non-GDPR websites both on all previously defined Section 4.1 covariates and on an additional pre-treatment outcome variable prior updates ($X_4$) that represents the number of times before GDPR adoption that a website updates its privacy policy:

$$\sum_{t'=X_0}^{2016} \text{update}(t)$$

where update is defined in Section 4.2 above. (Sample calculations for a website’s covariate vector ($\{X_{i,0}, X_{i,1}, X_{i,2}, X_{i,3}, X_{i,4}\}$) are included in Appendix B.2.) I discretize Rank ($X_3$) into distinct strata [1 , 1K), [1K , 10K), [10K , 100K), [100K , 1M), [1M $\infty$). I then matched GDPR and non-GDPR websites using a one-to-one, without-replacement, Mahalanobis matching algorithm as previously discussed in Section 3. The final matched set contains 4,900 GDPR websites spanning 48,020 privacy policies and 4,900 non-GDPR websites spanning 36,162 privacy policies. I use this matched GDPR and non-GDPR set to control for confounding bias and validate causal conclusions in all following analyses.
Table 4.1: Summary Statistics: GDPR and non-GDPR Covariates ($X$)

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treatment</th>
<th>Diff $\Delta_{raw}$</th>
<th>$t$-statistic$^{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Year ($X_0$)</td>
<td>2011.11</td>
<td>2013.40</td>
<td>2.29 (0.09)</td>
<td>54.68 ($&lt; 0.001$)</td>
</tr>
<tr>
<td>Duration ($X_1$)</td>
<td>5.48</td>
<td>4.45</td>
<td>1.03 (0.04)</td>
<td>25.21 ($&lt; 0.001$)</td>
</tr>
<tr>
<td>Category ($X_2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>49.25 %</td>
<td>61.57 %</td>
<td>12.33 %</td>
<td></td>
</tr>
<tr>
<td>Parked</td>
<td>0.23 %</td>
<td>0.23 %</td>
<td>4.63 %</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>0.23 %</td>
<td>0.23 %</td>
<td>3.07 %</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>0.23 %</td>
<td>0.23 %</td>
<td>1.42 %</td>
<td></td>
</tr>
<tr>
<td>Rank ($X_3$)</td>
<td>97,244.93</td>
<td>101,272.96</td>
<td>4,027.97</td>
<td>2.773 (0.005)</td>
</tr>
<tr>
<td>[1, 1K)</td>
<td>0.23 %</td>
<td>0.46 %</td>
<td>0.233 %</td>
<td></td>
</tr>
<tr>
<td>[1K, 10K)</td>
<td>2.32 %</td>
<td>1.740 %</td>
<td>1.740 %</td>
<td></td>
</tr>
<tr>
<td>[10K, 100K)</td>
<td>23.79 %</td>
<td>28.38 %</td>
<td>4.59 %</td>
<td></td>
</tr>
<tr>
<td>[100K, 1M)</td>
<td>26.63 %</td>
<td>29.09%</td>
<td>2.459 %</td>
<td></td>
</tr>
<tr>
<td>[1M, $\infty$)</td>
<td>47.02 %</td>
<td>38.00 %</td>
<td>9.026 %</td>
<td></td>
</tr>
<tr>
<td>Prior Update ($X_4$)</td>
<td>1.19</td>
<td>0.94</td>
<td>1.19 (0.32)</td>
<td>76.71 ($&lt; 0.001$)</td>
</tr>
</tbody>
</table>

Notes:
(1) $t$-statistic tests the null hypothesis $H_0 : \mu_c = \mu_t$ against the alternative hypothesis $H_A : \mu_c \neq \mu_t$.
(2) Normalized difference measures the controlled difference in sample averages as:

$$\Delta_{norm} = \frac{\bar{X}_t - \bar{X}_c}{\sqrt{\sigma_t^2 + \sigma_c^2}/2}$$  \hspace{1cm} (4.8)

where $\sigma_t^2$ and $\sigma_c^2$ represent the conditional within-group sample variances of the treatment and control, respectively.
4.5 Limitations

This chapter has three limitations. First, the Wayback Machine may have failed to retrieve older versions of archived homepages due to *robots.txt* blocks, non-standard markups, and password-protected logins, and the “privacy policy” dataset may contain less robust privacy policy data from earlier years. Second, I limited the CDX Server Wayback Machine crawl to English language homepages, but future research may include non-English languages and generate a more robust dataset of international privacy policies. Third, GDPR classification (Section 4.3) is severely limited and should be viewed as a starting point for additional research. GDPR classification based on regular expression matching likely failed to identify many GDPR websites (Matrix 4.1). GDPR privacy policies are not explicitly required to include a GDPR compliance term, and GDPR Article 12 may even *disincentivize* GDPR privacy policies from referencing legal jargon [20, 19].
Chapter 5

Quantifying GDPR Impact

This chapter addresses the fundamental question of whether the GDPR has actually changed privacy policy content. Section 5.1 uses a difference-in-difference estimator to quantify the causal impact of GDPR regulation on privacy policy change overall. The remaining three sections address nuances of GDPR impact on privacy policy change. Section 5.2 assesses whether the GDPR caused more websites to update their privacy policies through survival analysis. Section 5.3 analyzes whether the GDPR caused websites to update their privacy policies more frequently, and Section 5.4 assesses whether the GDPR caused websites to update their privacy policies more significantly.

5.1 Overall

This section addresses the most fundamental question: has the GDPR changed privacy policy content at all? I estimate the GDPR’s causal effect on approximate privacy policy FuzzyWuzzy similarities defined in Chapter 4. Figure 5.1 plots the period-to-period privacy policy similarities of GDPR and non-GDPR matched websites. I use a matched difference-in-difference estimator the first estimate of GDPR causal impact on privacy policy change and find that the GDPR has caused privacy policies to change 12.45% more than a non-GDPR control.
Overall, GDPR regulation appears to have significantly changed privacy policies. The next three sections explore how specifically the GDPR drove this privacy policy change. Did GDPR privacy policies change because *more websites* began updated their privacy policies? Did GDPR privacy policies change because the same websites began updating their privacy policies *more frequently*? Or did GDPR privacy policies change because websites began updating their policies *more significantly*?

### 5.2 Survival Analysis

In this section, I use non-parametric survival methods to analyze GDPR impact on the *number of websites* that update their privacy policies. A direct logistic model of GDPR and
non-GDPR update probability is severely limited by the short post-treatment observation period. Websites that have not updated their privacy policies yet cannot be assumed to never update. I instead model time until a privacy policy update and characterize websites with unreasonably large update times as never updating.

Let $T_k$ represent time from a website adopting a privacy policy to that website updating its privacy policy for the $k$th time. I first quantify the impact of GDPR regulation on time until a website’s first privacy policy update ($T_1$). (A histogram of $T_1$ distributions is shown in Figure 5.7.) The Kaplan-Meier survival curves shown in Figure 5.5b indicate that GDPR websites, on average, have a consistently shorter times until a first privacy policy update than non-GDPR websites. Figure 5.3b shows that all GDPR websites but not all non-GDPR websites are predicted to update their privacy policies at least once with a sufficiently long follow-up period. An Upper Asymptote Fitter through the Lifelines Python library quantifies the cure asymptote displayed in Figure 5.3b and estimates 25% of non-GDPR websites to never update their privacy policy.

I next consider GDPR impact on time until a second and third privacy policy updates ($T_2$ and $T_3$). (Histograms of $T_2$ and $T_3$ are shown in Figure 5.4.) The Kaplan-Meier asymptotes for GDPR $T_2$ and $T_3$ shown in Figure 5.5 indicate that GDPR websites are expected to update their privacy policies not just once after GDPR adoption but second and third times as well. An Upper Asymptote Fitter fit to non-GDPR websites cure asymptotes predicts that 37% of non-GDPR websites will never update their privacy policies a second time and 40% will never update a third time.

\footnote{Naively, only 14.79% of GDPR websites have not updated since GDPR adoption while 88.87% of non-GDPR websites have not updated their privacy policies since GDPR adoption.}
Figure 5.2: Histogram of Time until First Update: Distributions of $T_1$ for Matched GDPR and non-GDPR Websites

Figure 5.3: Estimated Survival and Hazard ($T_1$): Non-Parametric Estimates for Matched GDPR and non-GDPR Websites Survival and Hazard

Notes: GDPR Kaplan-Meier survival is convex, indicating that GDPR websites have a higher probability of first updating their privacy policy immediately after adopting it. Non-GDPR Kaplan-Meier survival is linear, indicating that non-GDPR websites have a constant update hazard over time.
Figure 5.4: Time until Second and Third Update Histogram: Distribution of $T_2$ and $T_3$ for Matched GDPR and non-GDPR Websites

Notes: The mean time to second update ($T_2$) is 5.645 periods for GDPR websites and 6.423 periods for non-GDPR websites. The mean time to third update ($T_3$) is 6.148 periods for GDPR websites and 7.32 periods for non-GDPR websites.

Figure 5.5: Estimated Survival and Hazard ($T_2$ and $T_3$): Non-Parametric Estimates for Matched GDPR and non-GDPR Websites Survival

5.3 Update Frequency

In this section, I next analyze the GDPR’s causal impact on the number of times a website updates its privacy policy after GDPR adoption. A plot of the percentage of GDPR and non-GDPR websites that update their privacy policies over time is shown in Figure
GDPR enforcement appears to be associated with a dramatic spike in GDPR privacy policy updates—51.54% of all GDPR websites updated their privacy policies in 2018B. Table 5.1 shows Wilcoxon ranked-sum tests across GDPR and non-GDPR groups pre-GDPR and post-GDPR. I again use a matched difference-in-difference estimator to calculate GDPR causal impact on privacy policy update frequency and find that the GDPR has caused GDPR privacy policies to update 25.43% more than a non-GDPR control.

Figure 5.6: Period-by-Period Percentage of Matched GDPR and non-GDPR Websites that Update

![Figure 5.6: Period-by-Period Percentage of Matched GDPR and non-GDPR Websites that Update](image-url)
Table 5.1: Wilcoxon Signed-Rank Test: Update Probability Comparison between Matched GDPR and non-GDPR Populations

<table>
<thead>
<tr>
<th></th>
<th>GDPR</th>
<th>non-GDPR</th>
<th>∆</th>
<th>t-statistic*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>µ₁ (σ²)</td>
<td>µ₀ (σ²)</td>
<td>(µ₁ − µ₀)</td>
<td></td>
</tr>
<tr>
<td>pre-GDPR</td>
<td>0.191 (0.108)</td>
<td>0.119 (0.082)</td>
<td>0.072</td>
<td>3.852 (&lt; 0.05)**</td>
</tr>
<tr>
<td>post-GDPR</td>
<td>0.399 (0.108)</td>
<td>0.145 (0.082)</td>
<td>0.254</td>
<td>42.301 (&lt; 0.001)</td>
</tr>
</tbody>
</table>

Δ (µ₁ − µ₀) = 0.207

Notes:
(1) All t-statistics are non-parametric Wilcoxon signed-rank t-statistics. Shapiro-Wilk Normality tests show neither GDPR nor non-GDPR period-to-period privacy policy changes are normally distributed.
(2) Matching appears imperfect as GDPR and non-GDPR websites show a small but significant difference in privacy policy updates pre-treatment even though our Matching algorithm explicitly controlled for Prior Updates.

5.4 Update Significance

Thus far, I have shown that the GDPR has caused more websites to update their privacy policies more frequently. However, perhaps GDPR websites are merely replacing what was one previously large update with two smaller updates? In this section, I lastly analyze GDPR causal impact on privacy policy update significance defined as the average FuzzyWuzzy similarity of only privacy policy updates (rather than the FuzzyWuzzy similarity of all privacy policies period-to-period which was analyzed in Section 5.1). Figure 5.7 plots average GDPR and non-GDPR update significance over time and shows minimal difference between GDPR and non-GDPR groups. Table 5.2 uses Wilcoxon signed-rank tests to compare update signif-
icance across matched GDPR and non-GDPR websites. A matched difference-in-difference estimator indicates that GDPR regulation has changed update significance a mere 5.63% more than a non-GDPR control.

Figure 5.7: Average Period-by-Period Update Significance for GDPR and non-GDPR Websites
Table 5.2: Wilcoxon Signed-Rank Test: Update Significance Comparison between Matched GDPR and non-GDPR Populations

<table>
<thead>
<tr>
<th></th>
<th>GDPR</th>
<th>non-GDPR</th>
<th>∆</th>
<th>t-statistic(^{(1)})</th>
<th>(p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-GDPR</td>
<td>65.431</td>
<td>67.641</td>
<td>-2.209</td>
<td>-1.212</td>
<td>(0.225)</td>
</tr>
<tr>
<td>post-GDPR</td>
<td>56.189</td>
<td>61.819</td>
<td>-5.630</td>
<td>3.511</td>
<td>(&lt; 0.001)</td>
</tr>
<tr>
<td>∆ (μ(<em>{post}) − μ(</em>{pre}))</td>
<td>-9.243</td>
<td>-5.822</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{(1)}\)All t-statistics are non-parametric Wilcoxon signed-rank t-statistics. Shapiro-Wilk Normality tests show neither GDPR nor non-GDPR period-to-period privacy policy changes are normally distributed.

5.5 Takeaways

This chapter proves that the GDPR caused a notable change in privacy policies. A matched difference-in-difference estimator shows that the GDPR has caused GDPR privacy policies to change 12.45% more than a non-GDPR control. In particular, the GDPR has incentivized an additional 25% of websites to update their privacy policies and caused websites to update 25.43% more frequently, ceteris paribus. The GDPR does not appear to greatly impact privacy policy update significance either positively or negatively. Chapter 6 analyzes whether the GDPR has positively changed privacy policy content.
5.6 Limitations

This analysis only considers FuzzyWuzzy similarity-based metrics and fails to account for changes in privacy policy readability, length, semantics, and HTML organization. Future research may consider alternative similarity metrics such as hierarchical clustering or WordNet path-length. Additional limitation related to causal assumptions are later discussed in Section 6.4.
Chapter 6

Qualifying GDPR Impact

This chapter addresses the follow-up question of how the GDPR has changed privacy policies. Section 6.1 analyzes GDPR impact on privacy policy length and readability. Section 6.2 analyzes GDPR impact on GDPR Article compliance.

6.1 Accessibility

This section analyzes the GDPR’s impact on privacy policy length ($Y_{\text{length}}$) and Flesch–Kincaid Grade Level readability ($Y_{\text{flesh}}$) metrics defined in Chapter 4. Figure 6.1a and Figure 6.1b plot average GDPR and non-GDPR privacy policies length and readability over time. Table 5.2 shows Wilcoxon signed-rank test results that compare GDPR and non-GDPR websites and shows GDPR privacy policies have become 67.85% and 0.36 grade levels less readable after GDPR adoption. Wilcoxon signed-rank $t$-statistics show significant underlying differences between GDPR and non-GDPR websites before GDPR adoption that may undermine causal conclusions.
Figure 6.1: **Accessibility Measures**: Average Length and Readability of GDPR and non-GDPR Privacy Policies

**(a) Policy Length (Y_{Length})**

**(b) Policy Readability (Y_{FKGL})**

*Notes:* GDPR privacy policies have a mean length of 18,481 words after GDPR adoption. GDPR privacy policies have a mean 13.34 Flesch–Kincaid Grade Level score after GDPR adoption, and over 76% exceed a *college* reading level.
### Table 6.1: Wilcoxon Signed-Rank Test ($Y_{length}$): Average Length Comparison of GDPR and non-GDPR Privacy Policies

<table>
<thead>
<tr>
<th></th>
<th>GDPR</th>
<th>non-GDPR</th>
<th>$\Delta$</th>
<th>$t$-statistic$^{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_1$</td>
<td>$\mu_0$</td>
<td>$(\mu_1 - \mu_0)$</td>
<td>$(p$-value)</td>
</tr>
<tr>
<td>pre-GDPR</td>
<td>11,010.11</td>
<td>6,096.52</td>
<td>4,913.61</td>
<td>7.31 ($&lt; 0.001$)$^{(2)}$</td>
</tr>
<tr>
<td>post-GDPR</td>
<td>18,481.47</td>
<td>7,367.93</td>
<td>11,113.57</td>
<td>4.66 ($&lt; 0.001$)</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>7,471.36</td>
<td>1,271.42</td>
<td>6,199.94</td>
<td>4.66 ($&lt; 0.001$)</td>
</tr>
</tbody>
</table>

**Notes:**

$^{(1)}$All $t$-statistics are non-parametric Wilcoxon signed-rank tests for $H_0 : \mu_1 = \mu_2$ and $H_0 : \mu_{pre} = \mu_{post}$ respectively.

$^{(2)}$GDPR and non-GDPR websites show statistically significant differences in both privacy policy length pre-GDPR ($p < 0.05$).

### Table 6.2: Wilcoxon Signed-Rank Test ($Y_{flesh}$): Average Readability Comparison of GDPR and non-GDPR Privacy Policies

<table>
<thead>
<tr>
<th></th>
<th>GDPR</th>
<th>non-GDPR</th>
<th>$\Delta$</th>
<th>$t$-statistic$^{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_1$</td>
<td>$\mu_0$</td>
<td>$(\mu_1 - \mu_0)$</td>
<td>$(p$-value)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-GDPR</td>
<td>13.25 (3.29)</td>
<td>11.92 (2.04)</td>
<td>1.33</td>
<td>3.77 ($&lt; 0.001$)$^{(2)}$</td>
</tr>
<tr>
<td>post-GDPR</td>
<td>13.61 (3.29)</td>
<td>11.87 (2.37)</td>
<td>1.74</td>
<td>4.52 ($&lt; 0.001$)</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>0.36</td>
<td>0.05</td>
<td>0.31</td>
<td>1.74 (0.093)</td>
</tr>
</tbody>
</table>

**Notes:**

$^*$All $t$-statistics are computed using non-parametric Wilcoxon signed-rank tests for $H_0 : \mu_1 = \mu_2$ and $H_0 : \mu_{pre} = \mu_{post}$ respectively.

$^{(2)}$GDPR and non-GDPR websites again show statistically significant differences in both privacy policy readability pre-GDPR ($p < 0.05$).
6.2 Compliance

This section analyzes GDPR privacy policies' compliance with three GDPR Article requirements: Article 27, Article 13, and Article 6.

**Article 27: Data Protection Officer Contact Information.** GDPR Article 27 mandates that all privacy policies include the contact details of an appointed Data Protection Officer (DPO). I assess whether more GDPR privacy policies include an email or address contact after GDPR adoption. I used regular expression matching and the spaCy Python library to identify contact patterns (i.e., named entities, address numbers, and emails) in privacy policies. The percentage of GDPR and non-GDPR privacy policies that include an email and an address over time are plotted in Figure 6.2a and Figure 6.2b. Naively, 7.6% more GDPR privacy policies appear to include a physical address and 15.2% appear to include an email address after the GDPR compared to a matched non-GDPR control. Additional analysis is needed to draw a causal conclusion.
Figure 6.2: DPO Contact: Percentage of GDPR Privacy Policies that Include Contact Details

(a) Includes Contact Address

(b) Includes Data Protection Officer Email

Article 13: Data Subject Rights  GDPR Article 13 requires all privacy policies to explicitly inform users of their “right to request access, rectify, erase, or restrict processing of personal data.” I next assess whether more GDPR privacy policies mention “data subject rights” after GDPR adoption. I identify various terms that crudely signal a mention of a data subject right (i.e., “access”, ”delete”, “portability”, etc.) based on prior research and
use regular expression matching to manually search GDPR privacy policies for mention of a keyword [18, 19]. Figure 6.3 shows the percentage of GDPR websites that mention various “data subject rights” in their privacy policies over time. Naively, 28% more GDPR privacy policies mention the “right to erase”, 35% more mention the “right to request access,” 16% more mention the “right to rectify,” and 23% more mention the “right to portability” after GDPR adoption. Future research may extend this analysis to non-GDPR control privacy policies to assess causality.

**Figure 6.3: Data Subject Rights:** Percentage of GDPR Privacy Policies that Mention Data Subject Rights

Article 6: Basis of Processing  GDPR Article 6 requires all personal data to be processed in accordance with one of six legal bases for processing: “Contractual Obligation”, “Legal Obligation”, “Vital Interest”, “Public Interest”, “Legitimate Interest”, and “Consent” (Appendix A.3). I lastly analyze whether more GDPR privacy policies mention a legal basis of processing after GDPR adoption. Figure 6.4 plots the percentage of GDPR privacy
policies that mention the six legal bases over time. Naively, 23% more GDPR privacy policies appear to include a legal processing basis immediately after GDPR enforcement in 2018B. GDPR regulation does appear to increase mentions of other processing bases: 45% more GDPR privacy policies mention “legitimate interest” and 28% more mention “contractual obligation” after GDPR enforcement.

**Figure 6.4: Legal Basis of Processing:** Percentage of GDPR Privacy Policies that Mention A Legal Basis of Processing

![Graph showing the percentage of GDPR privacy policies mentioning different legal bases over time.](image)

*Notes:* “Consent” is the most frequently mentioned legal basis of processing both before and after GDPR adoption.

### 6.3 Takeaways

This chapter identifies a clear tension between readability and compliance. Results indicate that the GDPR mandate for “concise and readable” privacy policies has largely failed: GDPR privacy policies have become 67.85% longer and 0.36 grade levels less readable since

---

1. Alternative legal processing bases have important potential to combat consent fatigue. [19]
GDPR adoption. Results also indicate that the GDPR has positively impacted privacy policy compliance with various GDPR Articles: 15% more GDPR privacy policies mention a Data Protection Officer’s contact information and 23% more mention a legal basis for personal data processing after GDPR adoption. Additional research is needed on how to balance this seemingly intractable trade-off between accessibility and transparency. Is it possible for privacy laws to incentivize privacy policies that are both accessible and compliant? If so, why is the GDPR failing to improve privacy policy accessibility in the status quo? If not, should privacy laws prioritize accessibility or compliance? Are longer and more comprehensive privacy policies or shorter and less comprehensive privacy policies more advantageous to users?

6.4 Limitations

This chapter has two specific limitations. First, I only consider Flesch-Kincaid Grade Level readability metrics in Section 6.1. Additional research may test alternative metrics such as SMOG Index, Gunning Fog Index, or Passive Voice Index. Second, Section 6.2 compliance analysis relies heavily on basic regular expression matching and has limited accuracy. Additional research may better quantify GDPR Article compliance using Polisis automated labels and structured querying \[19\].

More broadly, Chapters 5 and 6 together both pose two main limitations. First, the matching algorithm described in Chapter 4 may not fully control for all potential confounders. Wilcoxon signed-rank test statistics in Tables 5.1, Table 5.2, and Table 6.1 all show slight pre-treatment outcome differences between GDPR and non-GDPR websites. Additional research may better match GDPR and non-GDPR websites on a more comprehensive covariate vector to validate the “parallel trends” assumption \[22\]. More generally, additional research may consider alternative inverse probability weighting or doubly-robust causal estimators \[38\].
Second, this thesis exclusively analyzes the GDPR’s impact on privacy policies. However, GDPR regulations extend far beyond privacy policies as discussed in Appendix A.2. Additional research may assess the causal impacts of GDPR regulation on subject access requests, data breach notifications, or other data sources not considered.
Chapter 7

Conclusion

To my knowledge, this paper is the first to assess the causal impact of GDPR regulation on privacy policies. I overcome serious previous data limitations by building two datasets: a dataset of 31,843 websites classified as GDPR and non-GDPR with 88.5% accuracy and a large-scale, longitudinal dataset of 317,396 historical privacy policies from 2004 to 2019 corresponding to the 31,843 websites. Both datasets allow me to compare GDPR and non-GDPR websites’ privacy policies before and after GDPR adoption and establish causality.

My findings are twofold. I first prove that the GDPR has significantly changed privacy policies by causing more websites to update more frequently. The GDPR has caused GDPR privacy policies to change 12.45% more than a non-GDPR control, ceteris paribus. In particular, the GDPR has caused an additional 25% more websites to update privacy that would not have otherwise. The GDPR has also caused GDPR websites to update their privacy policies 25.43% more frequently than otherwise would have.

Second, I identify an inherent tension between accessibility and compliance in privacy policies. I find that GDPR regulation has made privacy more transparent but less readable: GDPR privacy policies have become more detailed to comply with various GDPR Articles but 67.85% longer and 0.36 grade levels less readable as a result. Additional research is needed on how privacy laws may account for this intractable accessibility-compliance trade-
off and incentivize privacy policies that are more readable.

The GDPR has become particularly relevant to recent policy discussions. Over past years, Brazil, South Korea, Argentina, New Zealand, India, South Africa, and California have all adopted privacy laws heavily modeled off of GDPR regulations without a clear understanding of GDPR successes and limitations [39]. My results show that the GDPR has had a clear impact on privacy policy content and many successive privacy developments have tremendous potential to strengthen digital privacy. However, my results also indicate that if the GDPR is to successfully serve as a model for future privacy laws, policymakers need to have a clearer understanding of the GDPR’s nuanced effects on privacy policies.
Appendix A: Legal

A.1 Market-Based Regulation

Legal scholars have extensively debated the merits of notice and choice [2, 10]. Advocates argue that “awareness and choice” benefits users by granting them greater control over their personal data. Users may sell or restrict the use of their data as they see fit. For example, suppose a user visits Amazon to buy a book. Perhaps that user does not care about privacy, or perhaps that user even likes that Amazon collects his personal data to tailor ads to his interests and autofill his billing information. That user can then “exchange” a bit of personal information for a valued Amazon book-buying service – reasonable quid pro quo. Suppose instead that a user cares deeply about his personal privacy and dislikes Amazon’s data collection practices. That user can then choose not to buy from Amazon and instead buy from an alternative site or not buy online at all. Advocates claim that an inherent tension often exists between their privacy concerns and their desire for “free” data-intensive services. “Awareness and choice” offers an appealing way to calibrate this tension.

Critics disagree and argue that market-regulation is a fundamentally broken approach to privacy. “Awareness and choice” depends not upon the laws of supply and demand but upon poor awareness and poor choice. First, privacy policies provide poor “awareness.” Users rarely understand the privacy policy conditions they are consenting to under ”awareness and choice” as discussed in Chapter 2. As a result, a discrepancy has emerged between users’ expressed privacy concern and marketplace behavior. Pew Research Center reports
that 79% of consumers are concerned for their privacy, yet the majority do little to protect
their personal data [40]. Only 28% of Internet users use a password protector, and only
19% use a private browser. Additionally, “awareness and choice” provide poor ”choice”. In
tivity, users are agreeing to a fair legal contract and free to opt-out. In practice, the price
of not engaging socially, commercially, and financially on the Internet is often too high for
consumers to truly opt-out. Notice and choice imposes a take-it-or-leave-it choice on users
that precludes genuine consent.

A.2 GDPR Overview

At 11 chapters and 99 articles, the GDPR is by far the most comprehensive privacy law in
global history. The GDPR extends far beyond privacy policy regulations discussed in this
thesis. A comprehensive chapter-by-chapter summary of all GDPR regulations is detailed
below [1, 6, 41].

Chapter I (General provisions) discusses the material and territorial scope of the GDPR.
Materially, the GDPR applies to all data controllers, data processors, and third parties
that process personal information. Territorially, the GDPR has a near-global reach. Article
3(1) (the “establishment criterion”) states that all controllers “established” in the European
Union are subject to GDPR regulation. Article 3(2)’ (the “targeting criterion”) further
extends GDPR regulation to any controller outside the EU that collects or processes the
personal information of a European resident. This extraterritorial scope extends even to the
United States. In 2015, the European Court of Justice replaced the Safe Harbor Agreement
with an EU-US Privacy Shield Framework that requires US firm to self-certify and comply
with GDPR requirements in order to process European data [5].

Chapter II (Principles) reviews personal data processing restrictions. Article 5 mandates
that “all personal data be processed lawfully and in a transparent manner; collected for
a specified, explicit, and legitimate purpose; and limited to what is necessary.” Article 6
defines six "specified, explicit, and legitimate" purposes for data processing that are further defined in Appendix A.3. Article 7 mandates that user consent must be “freely given, specific, informed, and can be withdrawn at any time.” Article 9 mandates that sensitive data collection requires additional "explicit opt-in consent". Note that GDPR privacy policies’ Article 6 compliance is further analyzed in Chapter 6.2.

Chapter III (Rights of the data subject) defines various data subject rights. In particular, Article 15 gives users the right to access their personal data; Article 16 the right to rectify their personal data; Article 17 the right to erase their personal data; and Article 18 the right to restrict the processing of their personal data. Again, note that GDPR privacy policy compliance with Chapter III is further analyzed in Chapter 6.2.

Chapter IV (Controller and processor) outlines many data security regulations. Article 25 mandates that companies that process personal data should “implement appropriate technical and organizational measures” to appoint a Data Protection Officers, conduct Data Protection Impact Assessments, and pseudonymize all personally identifiable information. Article 33 requires websites to notify a supervisory authority of all personal data breaches "within 72 hours and without undue delay."

All remaining chapters discuss legal nuances. Chapter V discusses transfers of personal data to third countries or international organizations under the GDPR. Chapter VI states that the GDPR will be locally enforced by European member state’s Data Protection Authorities (DPAs). Chapter VII appoints a European Data Protection Board as the lead authority to facilitate GDPR cooperation and consistency within the European Union. Chapter VIII discusses penalties: violations of the GDPR may result in fines up to 20 million Euros or 4% of the total annual company revenue. Lastly, Chapter IX, Chapter X, and Chapter XI discuss specific processing situations and final provisions.
A.3 Legal Bases of Processing

The GDPR specifies six valid legal bases for data processing analyzed in Chapter 6 and listed below.

- Contractual Obligation. Data processing is necessary to fulfill contractual obligations a data subject entered into.

- Legal Obligation. Data processing is necessary to comply with a legal obligation.

- Vital Interest. Data processing is necessary to protect a data subject’s vital interests.

- Public Interest. Data processing is necessary for a public interest.

- Legitimate Interest. Data processing is necessary for a controller’s legitimate interest. A legitimate interest could theoretically apply to any type of reasonable processing that does not violate “the interests or fundamental rights and freedoms of the data subject, in particular where the data subject is a child”. Potential legitimate interests include marketing, fraud prevention, intragroup transfers, or IT security.

- Consent. A data subject has consented to his or her data processing. Article 32 stipulates that “consent should be a freely given, specific, informed and unambiguous affirmative act.” ‘Affirmative act’ means data subjects must explicitly opt-in; websites can no longer rely on default settings or pre-ticked ‘I agree’ boxes as the basis for consent. ‘Freely given’ means data subjects must have genuine choice and not suffer any detriment if they refuse consent.
Appendix B: Technical

B.1 Wayback Machine CDX Server

Listing B.1: A sample Wayback CDX Server GET request for www.example.com

class Snapshot(dict):
    def __init__(self, urlkey=None, timestamp=None, original=None, mimetype=None,
                 statuscode=None, digest=None, length=None):
        super(Snapshot, self).__init__()
        self.urlkey = urlkey
        self.timestamp = timestamp
        self.original = original
        self.mimetype = mimetype
        self.statuscode = statuscode
        self.digest = digest
        self.length = length
        self.snapshot_url = http://web.archive.org/web/%s/%s/ % (timestamp,
                                                                 original)

url = http://web.archive.org/cdx/search/cdx?url=example.com&output=json&
     from=201905&to=201905&limit=10&filter=!statuscode:200&filter=!mimetype:text/
     html&collapse=timestamp:8
res = req.get(url)
snapshots = res.text.split('
')

for snapshot in snapshots:
    snapshot_items = snapshot.split(' ')

for snapshot in snapshots:
    snapshot_items = snapshot.split(' ')

if len(snapshot_items) == 7:
    snap = Snapshot(snapshot_items[0], snapshot_items[1], snapshot_items[2],
                     snapshot_items[3], snapshot_items[4], snapshot_items[5], snapshot_items[6])
    snapshot_list.append(snap)
B.2 Database Samples

Table B.1 shows a section of the “website” database. Table B.2 shows a section of the longitudinal “privacy policy database” queried for site_id = 3108.

<table>
<thead>
<tr>
<th>Site Id</th>
<th>First Year</th>
<th>Duration</th>
<th>Rank</th>
<th>Categories</th>
<th>GDPR</th>
<th>Prior Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>3108</td>
<td>2016</td>
<td>3</td>
<td>3598.57 'business'</td>
<td>1</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>2008</td>
<td>11</td>
<td>36058.42 'shopping'</td>
<td>1</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>720</td>
<td>2012</td>
<td>7</td>
<td>113375.15 'business'</td>
<td>0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>412</td>
<td>2004</td>
<td>15</td>
<td>49571.13 'travel'</td>
<td>0</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

...
Table B.2: “Privacy Policy” Database Sample

<table>
<thead>
<tr>
<th>Site ID</th>
<th>((3)) (Year, Phase)</th>
<th>((3)) Policy Text</th>
<th>((4)) Fuz</th>
<th>((4)) Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>3108</td>
<td>(2008, A)</td>
<td>“Your privacy is very important..”</td>
<td>NaN</td>
<td>0.0</td>
</tr>
<tr>
<td>3108</td>
<td>(2008, B)</td>
<td>“Your privacy is very important..”</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>…</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3108</td>
<td>(2014, B)</td>
<td>“Your privacy is very important..”</td>
<td>98</td>
<td>0.0</td>
</tr>
<tr>
<td>3108</td>
<td>(2015, A)</td>
<td>“We respect your privacy..”</td>
<td>53</td>
<td>1.0</td>
</tr>
<tr>
<td>3108</td>
<td>(2015, B)</td>
<td>“We respect your privacy..”</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td>3108</td>
<td>(2016, A)</td>
<td>“We respect your privacy..”</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td>3108</td>
<td>(2016, B)</td>
<td>“In an effort to protect..”</td>
<td>44</td>
<td>1.0</td>
</tr>
<tr>
<td>3108</td>
<td>(2017, A)</td>
<td>“In an effort to protect..”</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td>3108</td>
<td>(2017, B)</td>
<td>“In an effort to protect..”</td>
<td>98</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>…</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes:

(1) *Site Id* is the “website” dataset’s primary key and “privacy policy” datasets foreign key. (2) I discretize *Rank* before including it in Section 3.10 matching. I dummy encode *Rank* and one-hot encode *Category* and before including both in all regressions. (3) \((Year, Phase)\) represents a given period \(p\).

(4) *Policy Text* represents the website’s un-pre-processed privacy policy in period \(p\). *Fuz* represents the fuzzy similarity ratio of the website’s privacy policy between period \(p\) and the preceding period. *Update* represents whether the website’s privacy policy significantly changed in period \(p\).
B.3 Princeton-Leuven Longitudinal Privacy Policies Corpus

The Princeton-Leuven Longitudinal Privacy Policies Corpus is a large-scale database of over 1 million historic privacy policies from 1998 to 2019 from 108,499 distinct websites [32]. Ryan Amos and Gunes Acar, creators of the Princeton-Leuven Privacy Policies Corpus, developed a custom privacy policy crawler to detect all HTML links (<a> elements) from archived homepages and classify each as a privacy policy link using regular expression matching on certain terms (i.e., “privacy policy”, “privacy statement”, “privacy”, “privacy notice”, “cookie policy”, “your privacy”, “your privacy rights”) [16]. Amos and Acar then downloaded historical privacy policies from privacy policy links using the Wayback Machine CDX Server and removed misclassified non-privacy policies using a random forest classifier with 98% precision and 93% recall [20].

B.4 FuzzyWuzzy Distance

This section walks through a simple FuzzyWuzzy calculation between strings ”kitchen” and ”kitten”. The Levenshtein algorithm in Equation 4.2 sequentially recurses through the letters to calculate minimum edit distance. The first three letters (’k’, ’i’, ’t’) are the same, and \( \text{lev}('kitchen', 'kitten') \) becomes \( \text{lev}('chen', 'ten') \). A difference emerges at the fourth index (\( t \neq c \)). The Levenshtein algorithm now must consider whether a deletion, insertion, or substitution at the fourth index would produce the smallest overall edit distance. A deletion would delete the ’c’ of ‘chen’, add 1 to the minimum edit distance, and then call \( \text{lev}('hen', 'ten') \). An insertion would insert ’t’ into string ’chen’, adds 1, and then calls \( \text{lev}('chen', 'en') \). Lastly, a substitution would replace ’c’ with ’t’ in ’chen’, add 1 to the minimum edit distance, and call \( \text{lev}('hen', 'en') \). The Levenshtein algorithm continues until at least one string is empty (\(|a| = 0\) or \(|b| = 0\)), at which point the minimum edit distance becomes
the length of the non-empty string. Continuing recursively, the Levenshtein algorithm finds that the minimum edit distance is $lev('kitchen', 'kitten') = 2$, and the fuzzy similarity ratio is $fuz('kitchen', 'kitten') = 1 - \frac{2}{7} = 0.77$. 
References


[26] “AWS | Alexa Top Sites - Up-to-date lists of the top sites on the web.”


[31] R. Al-Rfou, “polyglot: Polyglot is a natural language pipeline that supports massive multilingual applications.”


[34] A. Swartz, “html2text: Turn HTML into equivalent Markdown-structured text.”


[37] “Country code top-level domain - ICANNWiki.”


[41] “Post-GDPR Developments on Data Protection and Privacy Regulations Around the World - Security Boulevard.”