

# Essays on Privacy

by

Dana Turjeman

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Business Administration)  
in The University of Michigan  
2021

Doctoral Committee:

Professor Fred M. Feinberg, Chair  
Associate Professor Elizabeth E. Bruch  
Associate Professor Nigel Melville  
Associate Professor A. Yeşim Orhun

Dana Turjeman

turji@umich.edu

ORCID iD: 0000-0003-1445-2983

© Dana Turjeman 2021

All Rights Reserved

This dissertation is dedicated to my family

## ACKNOWLEDGEMENTS

This work would not have been possible without the guidance of Fred Feinberg. From the first moment I stepped into his office, he made me feel I am capable of becoming a successful scholar. This was not an obvious form of support, given that I was already in advanced stages of my first pregnancy, and without much knowledge of statistical modeling. Despite all of the various commitments vying for Fred's time, he has always made time to meet, guide and support my learning and growth. Deep discussions on privacy, ethics, modeling and everything in between made this dissertation to be what it is today. An invitation to join him for a project became the core of this dissertation, an invitation to use a rich dataset became two complete essays presented here today.

But Fred taught me much more than statistical modeling and academic writing; he taught me how to be compassionate, kind, a good parent and an excellent listener. He taught me how to provide constructive feedback, guide gently, and how to come up with ideas from mere discussions on social media. He acted as an adviser, a counselor and, despite his belief he's not THAT old (he is) - a father figure I needed when I was away from my own (amazing) parents.

Fred enabled me to learn, to make mistakes, to try and fail, to work and rest, to flourish both professionally and personally with just-the-right amount of pressure and with lots of his commitments. Fred understood that I'm doubling as a mom and a scholar, and so I enjoyed not only Challah recipes, but also actual Challah's delivered to my doorstep after I gave birth (and in many other occasions), along

with dozens of shipments of Legos, toys, bikes and other untouched toys that were waiting in Fred and Carolyn's home (and thank you Ben) for my daughters to enjoy. Despite his remarkable experience and successful career, Fred makes every person he works with feel equally smart and capable. While being modest, Fred also knew to show confidence whenever I needed his expertise and back, pushing bureaucracy with diplomacy and wisdom.

Another mentor I am grateful for is Yesim Orhun. Throughout the years, she was there to teach me about econometrics, identification, causal inference and writing. I admire her broad knowledge and her amazing ability to explain, ask questions and find answers with outstanding, unlimited, tools of both empirical and experimental work. Merely hearing her asking questions during departmental seminars were lessons I gained from, but I was also fortunate enough to enjoy her close guidance and support. Yesim was there for mutual cries over the challenges in being a parent of young children, and was also there to push me and to teach me how to ask the right questions and to work relentlessly to answer them. She taught me how marketing scholars can drive to make the world a better place, driving for equality, equity, and positive business practices. She was there to inspire me that my work can, and should, have a voice and a clear call for action in terms of policy making and technological advances; ones that will benefit society.

This dissertation is relying on a plethora of resources and data that Elizabeth Bruch worked relentlessly to acquire, and she provided to me on a silver platter in a clean SQL format. Without borders or limitations (well, except for an important NDA), she gave me freedom to explore domains that are beyond the scope of what the data were intended to have been used for. These data were the basis of this dissertation, but also the basis for my deep interest in privacy. A combination of coincidence and opportunity-grabbing made us able to analyze the consequences of a data breach, and to develop methods for privacy preservation. But this also took

guts and effort on Elizabeth's behalf - making sure the company is aligned and understands the reasoning behind the need for even more data after the company has been breached.

I also owe a great amount of gratitude to Nigel Melville, who joined the committee and, with thoughtful and careful comments, helped me make this dissertation to include deeper analysis of privacy preserving mechanisms, interdisciplinary review and opened my eyes to endless resources and avenues for future research. With his kind words and positive attitudes, it was a joy having him in my committee.

Throughout the years, I was fortunate to be surrounded by friends, colleagues and mentors. Brilliant faculty members from Ross' Marketing department and elsewhere in Michigan were always there, adhering to the outstanding Michigan culture of kindness and wisdom, for a quick chat or deep discussions and feedback. Among them: Carolyn Yoon, Aradhna Krishna, Puneet Manchanda, Rajeev Batra, Rick Bagozzi, Scott Rick, Anocha Aribarg, Justin Huang, Eric Schwartz, S. Sriram, Katherine Burson, Florian Schaub, Sol Berman, and Inbal (Billy) Shani, who also doubled as a personal mentor and a big sister.

In addition to faculty, I was fortunate to meet friends who adopted me as I moved to Ann Arbor - bearing a new country, new language, new challenges, and a newborn. They all provided support and wisdom when I needed them most. More joined along the way, and were there to hear my complaints, to listen to my countless requests for jumping photos (and appreciate those later), and offer their shoulder to cry on, near the finish line. Despite the risk of missing some, I will mention explicitly Mike Pallazolo, Linda Hagen, Jenny Olson, Xu Zhang, Tong Guo, Tiffany Vu, Guy Shani, Rebecca Chae, Prashant Rajaram, Steve Shaw, Gwen Ahn, Tim Doering, Hayoung Cheon, Junyu Wang, Yunfei (Jesse) Yao, Bindan Zhang and of course - Longxiu Tian.

Beyond being a "big brother", Longxiu dared to take a project with me - a project I didn't even think would be theoretically possible - and made it happen. He pushed

this project with me, up the hill towards the finish line, holding his newborn baby in one hand and coding with the other. I am in deep gratitude for his friendship, belief in me and in our project, fun and productive collaboration, and endless hours of refinement on a project that we hope will enable privacy preserving data fusion and data protection. Importantly, this could not have been enabled without his co-pilot, Sharla. The strong couple, with their newborn, taught good-old-me a lesson or two on combining early parenthood with successful careers, and the PPDF project has benefited from their great wisdom, collaboration and strength.

In addition to this professional support nearby, from the other side of the world, my broader family supported me in endless ways. They were there for remote calls with parenting tips, countless babysitting hours and many celebrations they didn't quite follow, but played along. My parents educated me to dream big and to always seek challenges, but also encouraged me to skip school days in order to rest, rebalance, and enjoy life. This balanced way of living allowed me to run the marathon the PhD program is, and I am grateful for this and for their endless support.

Finally, this journey could not have even started with Erez. I moved to Ann Arbor with the best partner I could have ever dreamed of, and together we grew and developed. Our family doubled in size throughout my dissertation, and so Noga and Shavit were there to enjoy econometric classes, seminars, many travels and one busy mom. Erez, Noga and Shavit were there to hug me after a long day or a distant trip. They were there to challenge me in discussions, to (rightfully) request for attention and to improve my tent and Lego-building abilities. They were there to balance me, to teach me and to make me the happiest person on earth. Erez is the best partner, friend, father, husband and supporter. Discussions with Erez, while dropping the girls off at school, while cooking dinner, or under the stars during camping nights, drove some of the inspiration to the work you see here, but has also been an outstanding balance and solid rock for the turbulence the PhD journey included.

# TABLE OF CONTENTS

DEDICATION . . . . .	ii
ACKNOWLEDGEMENTS . . . . .	iii
LIST OF FIGURES . . . . .	x
LIST OF TABLES . . . . .	xi
ABSTRACT . . . . .	xii
CHAPTERS	
<b>I. Introduction . . . . .</b>	<b>1</b>
<b>II. When the Data Are Out: Measuring Behavioral Changes Fol-         lowing a Data Breach . . . . .</b>	<b>9</b>
2.1 Abstract . . . . .	9
2.2 Introduction . . . . .	10
2.2.1 Data Breaches . . . . .	12
2.2.2 Effects of Exogenous Shocks . . . . .	17
2.3 Data Description . . . . .	19
2.3.1 Website and Data Breach . . . . .	19
2.3.2 Behavioral Data . . . . .	20
2.3.3 User Profiles . . . . .	21
2.4 Temporal Causal Inference . . . . .	23
2.4.1 Measuring Effects of an Exogenous Shock . . . . .	23
2.4.2 Causal Inference Assumptions With TCI . . . . .	26
2.5 Temporal Causal Forests . . . . .	33
2.5.1 Comparing Control and Treatment Users . . . . .	33
2.5.2 Construction and Estimation – TCF . . . . .	35
2.5.3 Temporal Causal Forests . . . . .	39
2.5.4 Sources of Heterogeneity . . . . .	41



2.6	Results . . . . .	43
2.6.1	Average Treatment Effect . . . . .	43
2.6.2	Observed Sources of Heterogeneity . . . . .	45
2.7	Robustness Checks . . . . .	48
2.7.1	Diff-in-Diff Model . . . . .	49
2.7.2	Generalized Synthetic Control Group . . . . .	50
2.7.3	Bayesian Synthetic Control Method . . . . .	51
2.7.4	Falsification Test: Placebo Test . . . . .	52
2.7.5	Simulation Studies . . . . .	53
2.8	Conclusion and Future Directions . . . . .	55
2.9	Appendices . . . . .	58
2.9.1	A1: Heterogeneous Treatment Effects Estimates . . . . .	58
2.9.2	A2: Differences in Differences Estimates . . . . .	59
<b>III. Our Data Driven Future: Promise, Perils, and Prognoses . . . . .</b>		<b>60</b>
3.1	Abstract . . . . .	60
3.2	Introduction . . . . .	61
3.3	The Promise and Perils of Data Collection . . . . .	63
3.3.1	Shopping and Search Data . . . . .	63
3.3.2	Geolocation Data . . . . .	65
3.3.3	Health and Genetic Data . . . . .	66
3.3.4	Dating and Relationships . . . . .	67
3.3.5	Social Networks . . . . .	68
3.3.6	Data Breaches . . . . .	69
3.4	Reducing the Perils of Data Collection . . . . .	70
3.4.1	Privacy by Default . . . . .	71
3.4.2	Code Transparency . . . . .	73
3.4.3	Model Transparency and Interpretation . . . . .	74
3.4.4	Federated Learning . . . . .	75
3.4.5	Control Over Data . . . . .	76
3.4.6	Data Protection – An Ongoing Solution . . . . .	77
3.5	Enhancing The Promise of Data Collection . . . . .	77
3.6	Conclusion . . . . .	80
<b>IV. Privacy Preserving Data Fusion . . . . .</b>		<b>81</b>
4.1	Abstract . . . . .	81
4.2	Introduction . . . . .	82
4.3	PPDF Methodology . . . . .	84
4.3.1	Variational Autoencoders (VAEs) . . . . .	89
4.3.2	Bidirectional Transfer Learning . . . . .	94
4.3.3	Privacy Preservation Measures and Controls . . . . .	97
4.3.4	Handling Missing Data and Selection Bias . . . . .	101
4.4	Simulation Studies . . . . .	105

4.5	Proposed Application: Anonymous Survey and CRM Data . .	108
4.5.1	CRM Data . . . . .	110
4.5.2	Survey Responses . . . . .	112
4.6	Summary . . . . .	113
4.7	Appendices . . . . .	114
4.7.1	A1: Summary Statistics of Survey . . . . .	114
4.7.2	A2: Response Rates for a Sample of Survey Questions	118
<b>BIBLIOGRAPHY . . . . .</b>		<b>119</b>

## LIST OF FIGURES

### Figure

2.1	Percent of active users . . . . .	22
2.2	Illustration of Temporal Causal Inference . . . . .	25
2.3	Percent of active users per Control/Treatment group . . . . .	26
2.4	Percent of active users before and after the announcement . . . . .	27
2.5	Average percent of active users across all cohorts . . . . .	28
2.6	Comparison of control and treatment groups . . . . .	31
2.7	Average treatment effects . . . . .	44
2.8	Percent of active users per Control/Treatment group with estimates	45
2.9	Distributions of the heterogeneity in treatment effects . . . . .	46
2.10	Sources of heterogeneity . . . . .	47
2.11	Sources of heterogeneity with main effects . . . . .	48
2.12	Average percent of active users with TCF, BSCM and Gsynth . . . . .	52
2.13	Results of simulation studies . . . . .	54
3.1	Illustration of proposed changes. . . . .	72
4.1	Illustration of PPDF of two datasets . . . . .	87
4.2	Illustration of Variational Autoencoder of a single dataset . . . . .	89
4.3	Illustration of Normalizing Flow . . . . .	92
4.4	Illustration of a single VAE with Normalizing Flow . . . . .	93
4.5	Results of VAE with and without Normalizing Flow . . . . .	94
4.6	Detailed architecture of PPDF . . . . .	95
4.7	Simulation results on MNIST dataset . . . . .	106
4.8	Simulation results – varying $\delta$ . . . . .	107
4.9	Simulation results – varying tuning parameters . . . . .	109
4.10	Survey Summary Table #1 . . . . .	114
4.11	Survey Summary Table #2 . . . . .	115
4.12	Survey Summary Table #3 . . . . .	116
4.13	Survey Summary Table #4 . . . . .	117

## LIST OF TABLES

### Table

2.1	Temporal Causal Forests Mean and Standard Deviation . . . . .	43
2.2	Heterogeneous Treatment Effects Estimates . . . . .	58
2.3	Differences in Differences Estimates . . . . .	59
4.1	Results of data fusion – MNIST dataset . . . . .	108

## ABSTRACT

We create troves of data with nearly every step we take, every button we click, and every query we submit. These data can be used to cater to us with services that better align with our desires. They can help us locate restaurants matching our tastes, build up social networks with individuals sharing similar characteristics, find a soul mate or distant relative, attain financial goals, detect our health conditions, and potentially assist in developing individualized medicine. However, misuse of the data can induce us to buy things we don't need, offer us things that might harm our health, lead to an addiction, or even imprison us in the absence of wrongdoing. These data might also be breached, causing harm to us and our loved ones with revelations we might have never shared with the world.

In this dissertation, in a series of three chapters, I detect opportunities and propose approaches to reduce the potential risks and leverage the benefits of data collection and data usage. I first analyze users' reactions to the data breach in a matchmaking website, exploring their engagement changes and potentially insufficient behaviors in privacy protection following the breach. I then plot how years of data collection in the Marketing realm and other business domains have led to great improvements to our lives, but have also introduced harms – some of which are still likely awaiting revelation. I discuss potential avenues for improving the benefits of the vast data we all create, while reducing the risks associated with those data. Finally, I explicitly develop one of these solutions – a privacy preserving data fusion methodology – intended to securely combine datasets while reducing the risks of de-identification.

This dissertation, I hope, will serve as a steppingstone towards making the Marketing domain a safer zone in terms of privacy preservation. Marketing efforts were

a major driver towards vast data collection and the associated benefits and harms; the marketing domain can now drive the efforts to further improve the benefits and reduce those harms.

# CHAPTER I

## Introduction

Governments and firms collect troves of data on almost every activity we do. Over the course of years notable for outstanding technological improvements, these data have been used to enhance our efficiency, communication, and wellbeing. We are now able to find products with ease – even without searching for them – through recommendation systems that incorporate our purchase history and traits. We can determine the best nearby restaurant and navigate there efficiently. We can learn about potential health risks associated with our behaviors, genes, and food intake, from data we provide on our sleeping patterns, genome, and eating habits. We can connect to people in countless ways, locate long-lost school friends on social media, and even find a romantic partner based on quick survey of preferences. Most of our existential and social needs can be met with a click of a button. And all of these are direct products of the remarkable recent advances in data collection, data usage, marketing models, and machine learning, among other innovations.

However, with the great advantages of data collection also come the risks of invasions to privacy. These risks usually generate headlines in the “catchy” form of data breaches – when data are revealed to have been accessed by others who are not supposed to be able to. But in addition to these extreme breaches of trust and data, people are increasingly aware that invasions of privacy can also be felt when

presented with unsolicited offerings and ads that are seemingly out-of-context (Nissenbaum, 2009), have questionable resources of information (Kim et al., 2019), or have unclear or offending reasons to target us (Goldfarb and Tucker, 2011).

Even without the feeling of invasiveness, data collected on our daily activities can be used to cater to us with offerings that will unintentionally cause us financial and personal harm (Cowgill and Tucker, 2019; Lambrecht and Tucker, 2021), increase the likelihood for addiction, or deny us opportunities that others may receive. As an example, even basic health insurance can be out of reach for some, if health data are revealed without proper customer protection. Improper data collection and usage hurt our society well outside the marketing realm. They may be used in secretive or proprietary black-box algorithms to decide on imprisonment and potential societal risks (Simmons, 2017), and have been shown to reinforce societal biases (Kiritchenko and Mohammad, 2018; Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019; Dooley et al., 2021). But even if the risk of wrongful imprisonment or implicit biases are not major concerns (though they probably should be), other unknown risks from data usage, that cannot be accounted for by users when they are expected to provide “informed consent”, may cause harm. So complex are modern data infrastructures and agreements that, even companies that collect or use data can sometimes be unaware to what ends their data are eventually put to use – whether in their own proprietary black-box algorithms, or at the hands of a third-party.

Such reports on data misuse are increasingly making customers wary. Marketing efforts – previously focused on products’ abilities – are now shifting to protecting customers’ privacy. Companies such as Facebook, Google, Microsoft, and Apple, along with governmental entities such as the U.S. Census Bureau, are acknowledging the increased awareness and evident need for enhanced privacy protection. In recent years, extreme efforts have been made to align with regulations, customers’ expectations, and potential risks of data exposure. Marketing messages are shaped to inform



customers of these efforts. And slowly, marketing models themselves are beginning to align in order to preserve privacy.

In this dissertation, I discuss multiple methods that can be used by marketers, regulators and businesses to enhance privacy without forgoing the great benefits of data collection and usage. In Chapter II, joint work with Fred Feinberg, I explore heterogeneity in reactions to a massive data breach. We use data stemming from a matchmaking website, one for those seeking an extramarital affair, that was breached. The data, which were provided by the website to us under a nondisclosure agreement, include de-identified profiles of paying male users from the United States, and their activities on the website since joining, and up to 3 weeks after the disclosure of the data breach. A challenge in making causal inference in the setting of a massive and highly publicized data breach is that all users were informed of the breach at the same time. In such cases of “information shock,” there is no obvious control group. To resolve this problem, we propose Temporal Causal Inference. This procedure allows us to control for potential trends in both individual and temporal site usage, and allows us to extract insights regarding the homogeneous (average) treatment effect, along with nontrivial heterogeneity in responses to the data breach. We unsurprisingly find a significant decrease in user engagement on the website (in terms of searches and messages) and a significant increase in deletion of photos immediately after the announcement on data breach. However, the decrease in engagement shows signs of fairly quick attenuation – less than three weeks – after the data breach was announced. We further explore individuals’ heterogeneous reactions. Results show that users who perceive themselves as being more private on the website (by choosing not to share public photos, but rather only private ones) were the last to delete their photos after the breach, even though their photos had similar likelihood to have been breached. This false sense of privacy made those with high self-perceived privacy susceptible to greater risks.

Whereas in Chapter II I discuss data breaches, which can be seen as an obvious (and usually highly publicized) harm to privacy and brand trust, in Chapter III, also joint work with Fred Feinberg, I discuss the benefits and risks of data collection and data sharing. In addition to plotting explicitly the risks and benefits briefly mentioned above, we offer opportunities and propose solutions to mitigate the risks and increase the benefits – solutions that are especially suited for marketing efforts, but not only exclusively so. In this non-empirical piece, I develop an approach to allow businesses and individuals alike to avail of the vast opportunities possible using the tremendous amount of data we create each day. I discuss privacy-preserving methodologies and potential solutions to privacy risks, as well as avenues for further enhancements of the benefits of data collection. As I note in the chapter, some of the risks associated with data collection and data usage are yet to be revealed, as companies themselves don't always know how customers' data are used when making strategic or individual-level decisions. Chapter III, therefore, does not rely on merely “informed consent” of customers (though this is a necessary component): if companies are ill-informed of the consequences of the usage of customers' data, it's naive to expect such informed consent from the customers themselves. Instead of focusing solely on informed consent, the chapter focuses on solutions from multiple domains – behavioral sciences, law, computer science, economics, and (of course) marketing. These solutions range from opt-in for data collection, code and model transparency, federated learning, identity management tools with blockchain, and privacy-preserving methods for data publication, analysis, and fusion.

To exemplify one of my proposed ways of reducing the risks of data collection and data breaches, in Chapter IV, joint work with Longxiu Tian, I introduce a privacy-preserving method for data fusion. Data fusion – the act of combining multiple datasets – is a powerful technique to make inferences that are more accurate, generalizable, and useful than those made with any single dataset alone. However, when-

ever data fusion involves any form of user-level data, the technique poses a privacy hazard and increased risk of re-identification. Data fusion exercises in the realm of digital marketing and elsewhere should therefore be enacted with care so as not to reveal users’ identities as an unanticipated by-product of the fusion itself. We develop Privacy Preserving Data Fusion (PPDF), with the goal of preserving user anonymity while enabling the full suite of customer analytics allowed for by extant data fusion techniques. The PPDF framework consists of a set of variational autoencoders (VAEs) with bidirectional transfer learning (BTL). Conceptually, it builds upon advances in both Bayesian canonical correlation analysis and data matching autoencoders. Our framework does not require that the same users will appear in both datasets to make inferences on the joint data. It overcomes sample selection biases by recovering missing data in one dataset from additional variation in the other dataset. Moreover, PPDF is model-agnostic; that is, it allows for inferences to be made on the fused data, without the analyst needing to specify the model/analysis *a priori* to fusion. Most importantly, it does so while having Differential Privacy – state-of-the-art methodology for privacy preservation – built in, and without the original datasets ever coming in contact on a single machine or within a model, thereby reducing the risk of compromising users’ privacy and anonymity.

Prior approaches to privacy took one of two extreme approaches. The first is must-not-collect-data-in-the-first-place, and is exemplified by Shoshana Zuboff’s seminal work, in which the term “Surveillance Capitalism” was coined. Zuboff (2019) referred to most of the data collected today as “Behavioral Surplus”, and characterized acts of attracting people based on their data as a thinly-veiled gambit to make them waste money. The other approach is “privacy-is-dead” – privacy is long gone and you might as well enjoy the fruits the data bring. This approach is exemplified by “privacy nihilism” behavior – coined by Bogost (2018), and even before, by Scott McNealy (Sprenger, 1999), then CEO at Sun Microsystems, who suggested that “Privacy is

dead, deal with it.”

Evidence for customers’ behaviors in the spectrum of these approaches have been documented throughout the years, in a plethora of marketing, economics, and sociological essays. Customers state they care about privacy while clicking “I accept these terms and conditions” without ever reading them (Obar and Oeldorf-Hirsch, 2020) or while choosing the least private alternative (Athey et al., 2017). Some may show concern about targeted ads (Kim et al., 2019), while others might be more likely to click those ads if they have a sense of control over their privacy (Tucker, 2014).

In addition to varying customers’ behaviors, between the two extreme approaches of “collect no data” and “privacy nihilism” also reside a wide spectrum of prior work on technologies for privacy preservation. These technologies were developed primarily in the domains of information systems, cryptography, computer science, economics, and marketing. Over the years, myriad technologies have been developed in order to keep pace in the arms race against privacy attacks, whose sophistication evolved to circumvent existing safeguards.

Anonymization is typically the first step in hiding identities in datasets: simply removing identifiers eliminates the potential for detection, at least in theory. The definition of identifier, though, in itself evolved over the years, ranging from full names, social security numbers and home addresses, to email addresses, phone numbers and social networks information. However, anonymization of one identifier was seldom sufficient to effectively prevent individual-level identification. The combination of variables has also been used to reveal identities of so-called “anonymized datasets”. For example, the combination of 5-digit ZIP code, gender and date of birth might effectively allow a shadowy third party to identify more than half of the U.S. population within the U.S. Census data (Sweeney, 2000; Golle, 2006). Latanya Sweeney coined “quasi-identifiers” to acknowledge those combinations of variables that allow the unique identification of individuals. The prospect of machine learning methods

unleashed on vast databases would theoretically allow effective individual-level identification even with relatively little geodemographic information.

K-Anonymity was proposed by Sweeney (Sweeney, 2002) to assure that, with any combination of variables, at least  $K$  individuals in a database will share the same values, making definitive individual-level identification impossible. This can be practically accomplished by grouping levels of variables together (e.g., by storing age groups of five years instead of date of birth, or obscuring digits from a ZIP code), even endogenizing these procedures to account for the empirical nature of the data (e.g., larger bin widths where data are sparse, like for older members of a dating site). K-anonymity has been used for release and publication of datasets, and is still being used in password checkup tools such as “Have I been Pwned”, Google’s security checkup (Li et al., 2019), and in obscuring health data. However, it has been criticized for the ability to recover sensitive attributes if the  $K$  individuals who are sharing the same quasi-identifiers happened to have exactly the same values for sensitive attributes we did not want disclosed (in an attack referred to as “Homogeneity Attack”) (Domingo-Ferrer and Torra, 2008). In addition, when there is background information on the association between a sensitive attribute and the quasi-identifiers, such association may allow to determine or at least narrow the set of possible values of the sensitive attribute (Machanavajjhala et al., 2007), thus allowing for a “background information attack”.

As another step in the race against privacy attacks, L-Diversity and T-Closeness have been introduced (Li et al., 2007; Machanavajjhala et al., 2007). These methods improve on K-anonymity by assuring that individuals within a group will have enough diversity (at least  $L$  values, or having a wide enough distribution) in their sensitive attributes. At around the same time, “Differential Privacy” was introduced (Dwork et al., 2006b), and has been since refined to what is now considered the leading privacy preservation methodology. Differential privacy defines and rigorously limits

the chances of being identified as being in a dataset. With mathematical guarantees, and using added noise and randomization, differential privacy has been used for data publication, data synthesis, and now – with Chapter IV introduced here – it will allow for privacy-preserving data fusion.

In my work, I embrace the data revolution while acknowledging its risks and finding ways to mitigate or even eliminate them. People are actively seeking services and products better tailored to their desires – services that can assist them in attaining their goals. Taking the extreme only-privacy approach would mean forgoing all these great advances. At the same time, I argue that there’s no need to forgo privacy completely. We should strive to preserve privacy while still enjoying the full suite of data-enabled wisdom in our data-driven world.

## CHAPTER II

# When the Data Are Out: Measuring Behavioral Changes Following a Data Breach

### 2.1 Abstract

As the quantity and value of data increase, so do the severity of data breaches and customer privacy invasions. While firms typically publicize their post-breach protective actions, little is known about the social, behavioral, and economic aftereffects of major breaches. Specifically, do individual customers alter their interactions with the firm, or do they continue with “business as usual”? We address this general issue via data stemming from a matchmaking website, one for those seeking an extramarital affair, that was breached. The data include de-identified profiles of paying male users from the United States, and their activities on the website since joining, and up to 3 weeks after, the disclosure of the data breach. A challenge in making causal inference(s) in the setting of a massive and highly publicized data breach is that all users were informed of the breach at the same time. In such cases of “information shock”, there is no obvious control group. To resolve this problem, we propose *Temporal Causal Inference*: for each group of users who joined in a specific time period, we create an appropriate control group from all users who had joined prior to it. This procedure helps control for, among other elements, potential trends in both individual

and temporal site usage that broadly fall under the rubric of “normal” usage trajectories. Following the construction of suitable control groups, we apply and extend several causal inference approaches. We adapt Causal Forests (Athey et al. (2019), among other forest-based methods) into Temporal Causal Forests, to better align ‘temporal’ inference settings. The combination of Temporal Causal Inference and Temporal Causal Forests methods allows us to extract insights regarding the homogeneous (average) treatment effect, along with nontrivial heterogeneity in responses to the data breach. Our analyses reveal that there is a decrease in the probability of being active in searching or messaging on the website, and a notable increase in the probability of deleting photos, ostensibly to avoid personal identification. We investigate several potential sources of heterogeneity in response to the breach announcement, and conclude with a discussion of both managerial consequences and policy considerations.

## 2.2 Introduction

Seventy million customer accounts at Target were compromised in 2013, more than twice as many eBay accounts were breached in 2014 (eMarketer, 2014), and more than twenty million records of Uber’s passengers and drivers were breached in 2016<sup>1</sup>. The number of records revealed to be compromised in data breaches has increased dramatically over the past several years: in 2012, some 20 million records were compromised; in 2015 the number rose to 318 million, and in 2017 the number reached nearly 2 billion (a figure reduced to “only” 1.37 billion records in 2018)<sup>2</sup>.

---

<sup>1</sup>[www.fortune.com/2018/04/12/uber-data-breach-security](http://www.fortune.com/2018/04/12/uber-data-breach-security)

<sup>2</sup>[www.privacyrights.org](http://www.privacyrights.org) measures, presented here, include data records that were compromised due to security breaches. Possible causes are unintended disclosure, hacking or malware, and physical loss (both electronic, non-electronic, and stationary devices). All compromised records were from businesses, educational institutions, government and military, healthcare providers and nonprofit organizations. In reality, the number should be considerably larger; for many of the breaches listed, the number of records is unknown. This list is not intended as a comprehensive compilation of all breached data. For additional explanation regarding the associated measures, see [#](http://www.privacyrights.org/data-breach-FAQ)2



Despite these severe incidences, little is presently known about users’ reactions in the wake of a publicly disclosed data breach.

As with any “exogenous” information shocks – those that neither companies nor their customers can anticipate – disclosures of data breaches may result in heterogeneous reactions among users and customers<sup>3</sup>. Such varying responses can arise in several ways: experiments show that customers may vary in their perceptions of privacy and in the risk associated with data breaches (Athey et al., 2017); surveys suggest that customers may vary in their expectations of the company’s actions, reactions, and obligations both before and after the disclosure (Madden and Rainie, 2015); and they may vary in their engagement with the company, thereby needing its services more, or less, than other customers (Janakiraman et al., 2018). Assessing these sources of heterogeneity, and the range of reactions to a disclosure of a data breach (or any exogenous shock) is critical for firms, their customers, and the policymakers enacting guidelines to minimize potential damage. However, despite the importance of such measures, they are not easy to enact; in highly publicized shocks, it’s uncommon to have a group of users who remained uninformed, and can thereby serve as controls.

We develop and present a methodology for measuring changes in customers’ behavior following the public disclosure of an exogenous shock (an event that was not anticipated by either the customers or the company), applying it to a severe data breach that received worldwide media attention. This attention owed in large part to the nature of the focal website, one primarily intended for those seeking extramarital affairs. Thus, the breached, disclosed data included especially sensitive personal profiles, exposing users’ desire to engage in a relationship outside their primary one (at least for most users), as well as personally identifying information, such as credit

---

<sup>3</sup>The terms “customer” and “user” are used interchangeably throughout, except where ambiguity might arise.

card numbers, for all paying members<sup>4</sup>.

Our results suggest that relative to the appropriate counterfactual – which is enabled by the proposed Temporal Causal Forests methodology – users were less likely to engage in searches and messages on the website immediately after the data breach was announced, and they were significantly more likely to delete their photos than they otherwise would have. We also explore differences in users’ reactions to the breach, as well as potential reasons for this heterogeneity. In addition, by the third week after the breach announcement, there was an attenuation of some of these effects. We will discuss differences in privacy preferences that can be inferred from our analysis, and that may be useful in determining both effective privacy regulation and guidelines for companies’ reactions in future breaches.

The remainder of the paper is organized as follows. Section 2.2.1 and 2.2.2 reviews the literature on the effects of data breaches from the perspectives of the individual and of the company, and the challenges to measure any such exogenous shock to company’s perception. Section 2.3 describes our data, while Section 2.4 outlines the construction of control groups via *Temporal Causal Inference*. Section 2.5 develops *Temporal Causal Forests*, a non-parametric approach to assessing individual treatment effects using Causal Forests with Local Linear Correction. Section 2.6 details the results of these analyses, Section 2.7 provides various robustness checks, while Section 2.8 closes by discussing the results and the methods used, and avenues for future work.

### 2.2.1 Literature Review – Data Breaches

The Generalized Data Protection Regulation of the European Union (The European Union, 2016-05-04), as well as Security Breach Notification Laws, devised by

---

<sup>4</sup>Between 2014 and 2018, several websites, intended for the same purpose of extramarital affairs, have announced they suffered a data breach. Due to confidentiality, we do not disclose the name of the focal website, as well as the exact time during which the announcement took place. Some other technical details regarding the nature of the announcement are also removed for confidentiality.

all US States (of State Legislatures, 2018), were put in response to the spate of data breaches in recent years (among other reasons). Such laws require all governmental or private entities to disclose instances of data breaches as soon as these are brought to their attention, even if the breached data were not made public. These laws aim to reduce identity theft, mainly financial misappropriation in which criminal entities use personally identifiable information in order to adopt the identity of another person<sup>5</sup>. While lab studies have shown that data breach notifications are often neither clear nor particularly alarming to those whose data were breached (Zou and Schaub, 2019), Romanosky et al. (2011) found that such data breach notices successfully reduced the number of identity thefts caused by data breaches by 6.1%. In this study, we aim to develop a method to measure the consequences of such data breach notification, on the behavior of users/customers, of the affected company.

To the individual affected, data breaches may cause more than financial losses. They are perceived as privacy invasions, leading to lack of trust and potentially information leaks of many sorts: purchase behavior, daily routines, email correspondences, etc., along with identifiable information, all typically construed as “private”. Taylor (2004), in a survey, found substantial heterogeneity in customers’ preferences for privacy: some prefer to disclose their personal information and purchase behavior in order to gain lower prices and more accurate product suggestions, while others prefer to protect their anonymity by not disclosing such information. Acquisti and Varian (2005) confirmed the economic effects of individuals’ ability to protect their privacy, and heterogeneity in users’ preferences to remain anonymous, in the context of price discrimination based on past purchases that were monitored by the company. They

---

<sup>5</sup>To illustrate the extent and magnitude of financial outcomes of identity thefts, an estimated 17.6 million persons, or 7% of all U.S. residents age 16 or older, were victims of one or more incidents of identity theft, which resulted in cumulative loss of more than \$15.4 billion, in 2014 alone (Harrell et al., 2015). Notifications of data breaches encourage persons whose data were compromised to seek a remedy and protection of their financial identity through the use of identity theft tools. According to a survey reported by Ablon et al. (2016), 62% of respondents who had their data compromised following a breach accepted free credit monitoring offered by the company whose data were breached.

are challenged to find an optimal strategy, partially because the sources of heterogeneity in the willingness to disclose information are not clear. Athey et al. (2017) confirm, in a randomized experiment, heterogeneity in participants' willingness to disclose private information, such as the contact information of friends, for small financial incentives. Goldfarb and Tucker (2012) found that younger people tend to be more private, in the context of revealing their income.

Such heterogeneity among users can be found not only in the willingness to share information before any breach was associated with the company, but also following notification of a breach. One might surmise that companies whose data were breached would be deemed unworthy of continuing customer trust; yet, according to a survey by Ablon et al. (2016), 89% of respondents continued to conduct business with a breached firm, while 11% stopped cold. One percent of respondents reported *increasing* the amount of business they conducted with the breached firm, although this cannot be separated from ordinary engagement trajectories, a topic we return to for causal inference purposes. Immediately following Facebook's allegations of privacy misconduct in the infamous "Cambridge Analytica" case, a Reuters/IPSOS survey (May 2018) found that about half of Facebook's American users said they had not changed the amount that they used the site, and another quarter said they were using it more. The remaining quarter claimed they were using it less, stopped using it, or even deleted their account.<sup>6</sup> However, these measures were based on surveys, and customers may fail to state their actual behavior accurately.

Notification of data breaches results in direct and indirect financial outcomes to the firms whose data are compromised. Acquisti et al. (2006), Choong et al. (2016), and Rosati et al. (2017) showed that, following disclosures of data breaches, there are short-duration reductions in the company's stock market valuation, based on both the

---

<sup>6</sup>"Three-quarters Facebook users as active or more since privacy scandal: Reuters / Ipsos poll", May 2018: [www.reuters.com/article/us-facebook-privacy-poll/three-quarters-facebook-users-as-active-or-more-since-privacy-scandal-reuters-ipsos-poll-idUSKBN117081](http://www.reuters.com/article/us-facebook-privacy-poll/three-quarters-facebook-users-as-active-or-more-since-privacy-scandal-reuters-ipsos-poll-idUSKBN117081)

bid-ask spread and trading volume. Others, such as Gordon et al. (2011), agreed that there is a short-term reduction in stock market valuation, but measured a decrease in the magnitude of the effects of such breaches over the years: the authors suggest that familiarity with breaches, and the uptick in their frequency of occurrence, might diminish the negative impact of a breach on customers' loyalty to the firm. Amir et al. (2018) question the small reductions in market valuation and suggest that under-reporting of severe cases moderate outcomes. Extrapolating from data breaches that were revealed by the hackers and other entities outside the publicly-traded companies, they find a long-duration negative effect of data breaches on market value. Following a data breach, a harm in trust has been similarly documented. A survey conducted by Pew Research Center (Madden and Rainie, 2015) suggests that online service providers are among the least trusted entities when it comes to keeping information private and secure. When asked about search engine providers, online video sites, social media sites and online advertisers, the majority felt "not too confident" or "not at all confident" that these entities could protect their data. In a separate question, few respondents have reported to have "a lot" of control over the information that is being collected by such firms. These findings also suggest substantial heterogeneity in reaction to privacy violations.

Firms whose data were breached can suffer financially through loss of revenue, but also via punitive measures like monetary fines (Romanosky et al., 2014). Yet firms can mitigate or even reverse reputational damage through their reaction in the breach's immediate wake: most respondents highly value prompt notification, to the extent that there is sometimes an *increase* in valuation following a breach (Ablon et al., 2016). Publicly "shaming" companies using various media will only be useful if the company did not disclose the breach of its own volition and did not take appropriate precautions both before and following the breach itself. Some preliminary evidence suggests that firms can *benefit* from negative buzz; for example, Han et al. (2020)

found that, in some cases, negative attitudes towards a product or company may result in increased awareness and downstream purchase intent.

Kude et al. (2017) measured the ability to restore customers' sentiment (i.e., overall feeling towards the firm) after a breach: following the data breach to Target, which occurred in 2012 and affected 70 million customers, the authors surveyed 212 customers whose data were compromised. They found substantial differences in the perception of compensation offered by the company, and that these perceptions varied widely among respondents according to their personal traits. Zhong and Schweidel (2020) conducted an analysis of twitter conversations and found that most referencing the "Under Armor" Data breach were initiated by people who'd never discussed the brand before, suggesting that the brand might be more frequently discussed in crowds that had not previously been involved in such discussions. Moreover, the authors find that the negative sentiment following the data breach was short-lived. Taking the outcomes of Han et al. (2020) and Zhong and Schweidel (2020) together might suggest that the longer-term outcomes of a data breach announcement could yield a net gain in brand awareness, that is, a positive for the focal firm. Note that the examples above – of both the negative effects and ways to mitigate them – are based on stock market valuations, surveys, twitter sentiment or lab studies. One of the only documented empirical measures to a change in behavior of customers following a data breach, prior to this paper, is presented in Janakiraman et al. (2018): they suggest that heterogeneity in individual response should be considered when assessing the effects of a disclosure of a data breach. That is, customers may well have varying reactions, depending on, amongst other elements, the sensitivity of the data that were leaked, general level of concern about disclosure, and prior expectations in regard to the firm's safeguarding their personal data. Possible reasons for the lack of empirical measures will be discussed in section 2.2.2.

In summary, the literature on the effects of data breaches suggests that users will

differ in their reaction – with some even seeming to increase interaction with the breached company. Moreover, such differences can vary based on personal circumstances and perceived actions taken by the firm. While it is of obvious importance for firms to comprehend and anticipate such reactions, what is lacking is a suite of methods to disentangle typical individual-level usage trajectories from downstream *post hoc* behavior when essentially all users were “treated”, that is, made aware of the breach.

### **2.2.2 Literature Review – Measuring the Effects of Exogenous Shocks**

Firms face constant changes in the way users perceive them. Some of these changes are governed by the firm and can be either beneficial or detrimental to the customer: product enhancements and price increases provide, respectively, two such examples. Other changes are exogenous, unanticipated by the company, and may be caused by a criminal act, legislation, natural disasters, or an unexpected flaw in production, among other reasons.

Understanding the range of reactions to an exogenous shock, including data breaches, is a quantitative problem that merits methodological and substantive attention. It is possible to measure reactions to changes under the company’s full control, using A/B testing (randomized controlled trials), test markets, or other means. But, following a highly-publicized exogenous shock, it is difficult or impossible to identify a group of users who remained unaware of it, and so can serve as a control group for measurement purposes; moreover, users who somehow managed not to be informed of a major shock (such as a product recall, data breach etc.), cannot be viewed as representative of the larger pool they would be intended to represent. Lack of such control groups makes it difficult to evaluate and measure the consequences of the shock (Cleeren et al., 2017).

In addition, observable behavioral changes can arise for many reasons, irrespective

of the data breach notification or other exogenous shock. Therefore, when aiming to measure the causal effect of such a shock on individual behavior, it is important to measure it in comparison to the behavior that the users *would likely have engaged in*, had the announcement of the shock not been made, while accounting for heterogeneity both in the users' activities, and in their reactions to the shock.

Several solutions have been proposed in the Marketing, Accounting, and Economics literatures: In the context of data breaches specifically, Janakiraman et al. (2018), discussed earlier, compared the changes in sales over time in breached vs. non-breached channels. Measures of change in stock market value were also presented due to the availability of such data, for public companies (Acquisti and Varian, 2005; Amir et al., 2018). However, these measures assume consistent behavior of the corpus of customers among channels, assume no spillover effects between channels, or require that the breached company be public. As mentioned earlier, other methods relied on surveys or lab studies, and their generalization to real data breach notifications is limited. In the context of other exogenous shocks, such as product recalls and product-harm crises, Cleeren et al. (2017) note several empirical analyses, along with surveys and lab studies. Of the methods presented in this review paper, measures of aggregate change to sales, compared to other brands, were proposed, as well as financial event study methods (again, relevant mainly for public companies). To reiterate, all these methods either assume that there are comparable products, assume consistent behavior and no spillover effects, or provide only aggregate results.

Our overarching goal is therefore to estimate the heterogeneous treatment effects of an exogenous shock, on users that were already members of the focal site, while also acknowledging that they could well have changed their behavior even had the shock not occurred, due for example to typical or predictable behavioral trajectories. Our method is applied here to address the substantive question of the effects of a data breach, but is applicable to other contexts, so long as individual-level data are



available both before and after the shock.

In the following sections, we will describe the data to be used in our setting, introduce our identification strategy – *Temporal Causal Inference* – which allows the analyst to overcome the key stumbling block of lacking a “non-informed” control group, and adapt a set of non-parametric causal inference methods into *Temporal Causal Forests*, in order to assess the heterogeneous treatment effects attributed to the announcement of the data breach.

## 2.3 Data Description

### 2.3.1 Website and Data Breach

The focal data come from a matchmaking website aimed at those seeking extra-marital affairs, either online or in-person. The website suffered a massive data breach, which was announced in a manner ensuring widespread attention: unauthorized parties (henceforth, “hackers”) declared that they had downloaded detailed personal information of all users of the website. This personal information included email addresses, credit card information, preferences for affair types, among other potentially identifying and/or socially embarrassing elements. The announcement was highly publicized, and reached major media outlets in the US and abroad. In addition, the website made several announcements to their users and to the general public. Therefore, it is reasonable to assume that news of the breach reached the entirety of the web site’s user base in short period of time<sup>7</sup>.

It is important to note that we are not using any of the hacked data; rather, the data we use were provided directly by the firm, and conforms to a non-disclosure agreement. Our collaboration with the website afforded detailed user behavior records, as

---

<sup>7</sup>We also verified this assumption, to the extent possible, with detailed analyses of the various media publication dates, range thereof, and Google Trends around the name of the website, which spiked less than a day after the announcement.

well as de-identified profiles of all paying members. Our data window commences approximately two years before the announcement of the breach, through ~3 weeks of activity following the announcement. Importantly, the leaked data were not made public during this time, so we can treat the announcement of the data breach as a single exogenous shock unrelated to the aftereffects of publication of the data themselves.

### 2.3.2 Behavioral Data

For purposes of consistent reference, “behavioral data” pertains to activities taken on the focal site itself, including searches made, messages sent by the users, as well as such deliberative actions as deletion of photos. Our sample consists of all paying male users from the United States who had joined the website 2-6 months before the breach was announced, and had at least one “activity” (searches, messages, etc. – anything beyond the mere creation of the profile) on the website before the breach was announced. The relatively long span of join dates allows for an account of regularities in activity patterns *before* the breach, e.g., satiation, attrition, and/or other trends and fluctuations in individual-level behavior. Each user is assigned a unique ID that does not change over time, allowing us to view all the activities users made throughout the data window.

The focal website is a so-called “freemium” site, where one can join for free and enjoy limited functionality. A feature on this dating site is that women obtain almost “full” access for free, while men must pay a monthly access fee. Only such “full” members –can contact other members, but any user on the website can browse anyone else’s content, except for, in some cases that will be described shortly, other users’ photos. Due to this feature of the website, and the nature of the breached data, our analysis and statistical estimates pertain to male users that had paid for membership, and all such users had to provide their full name and home address in order to process

payment; this is not so for nonpaying users, many of whom were pseudonymous, and so we limit our purview to legitimate, accurately recorded male users. Consequently, these users were informed, on the day of the announcement of the breach, that their real names and addresses were in the hands of the hackers – entailing the risk of widespread exposure – along with other personal information, and an indication that they were seeking an affair.

The sample used for analysis consists of ~57K users, apportioned into 24 weekly cohorts (groups of users who join in the same week), based on week of joining. Figure 2.1 illustrates the percent of active users per cohort per week throughout the data span. As can be seen, the average number of active users prior to the breach initially increases and then gradually decreases. The breach happened 27 weeks from when the first cohort joined and, as stated previously, the data window extends to 3 full weeks after the breach announcement – and before the data were made public. From observation, the average number of users who deleted photos increased immediately following the announcement. However, for the average number of users who made at least one search or sent at least one message each week, the effect is not clear, due to the inconsistent number of activities prior to the breach, and to the natural decrease that would have presumably occurred even without the breach, rendering it difficult to determine whether the breach affected users’ behavior or not.

### **2.3.3 User Profiles**

Upon joining the website, users provide a full profile, which includes gender, marital status, date of birth, height, weight, ethnicity and many other geodemographic covariates. The specific covariates used in the analysis are marital status (attached

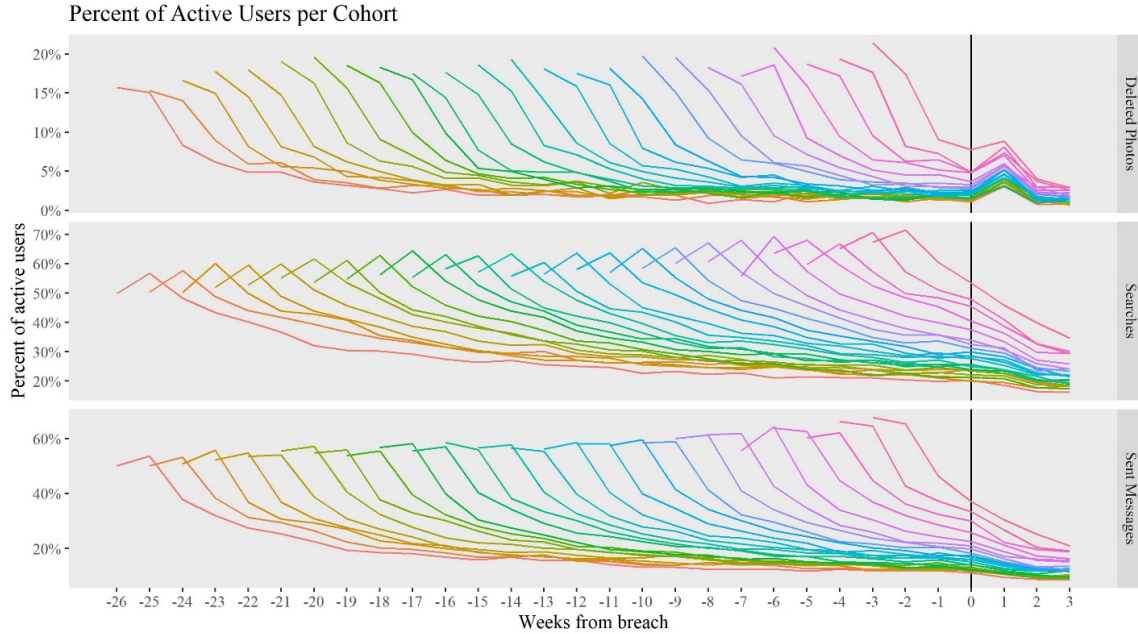


Figure 2.1: Percent of active users, for each cohort, as a function of the number of weeks relative to the announcement of the data breach. Different colors indicate week that cohort joined the website. Vertical lines indicate the announcement of the data breach.

or single<sup>8</sup>, where 67% of users are attached, and 33% single)<sup>9</sup>, and a binary indicator of whether the user’s public profile includes a photo. Users could choose whether any user on the website can see their photos, or only users that shared their photos reciprocally. Of all users, 85% had public photos, and 15%, referred to as “private”

<sup>8</sup>Marital status is declared by these (categorical) statements: (a) Attached male seeking female; (b) Single male seeking female, (c) Male seeking male. Only about 0.1% of the users in the estimation had status (c). Since (c) does not clearly state marital status, we coded (b) and (c) together as “single”.

<sup>9</sup>One possible concern is our ability to know whether the users are honest about their marital status on their profile, and whether they are in an “open relationship”. To verify both, we complement our data with a survey we conducted on a random sample of the website’s users, in an anonymous setting, with a declaration of academic objectives, long before and independently of the announcement of the breach. The survey results show similar percentage of married users in the survey and in the website (62% married, 7% living with a partner, and 3% in “serious relationship” in the survey, and 67% identify themselves to be “attached” in their profile on the website). This means that there is no clear bias in the disclosing of one’s marital status on the website. Even if there was a bias, the disclosure of one particular marital status is the effect we are interested in, and therefore, revealed status is the variable we use. In order to test for “open relationship”, we again refer to the survey. Of those who stated themselves to be in a committed relationship, only 7% said their spouse knows they were on the website. This allows us to largely rule out possible “open relationships” or other forms of socially acceptable affair-seeking among attached men.

users, required reciprocity in sharing all of their photos. These covariates would be expected to correlate with an important latent element of public disclosure: that some users had “more to lose” than others. Note that, because the user can change his profile (be it due to real changes in his life, changes to his privacy preferences, or other reasons), the covariates included in the analysis are potentially time-varying. To maintain consistency in such covariates such as marital status, we use only the user’s last profile prior to the breach. This is critical for a reason beyond mere consistency: users necessarily presumed that the hackers had this final profile and therefore it is this specific data that could potentially have gone public.

## 2.4 Temporal Causal Inference

### 2.4.1 Measuring Effects of an Exogenous Shock

Our goal is to estimate the effect of the data breach on the probability a user will be active on the website. We refer to the announcement of the breach as a single, exogenous *treatment*, for which we want to estimate the effect. As detailed earlier, the main challenge is that there is no clear *control* group, since it is reasonable to assume (due to the high publicity of the event and the nature of the public announcement) that all users were informed of the data breach at essentially the same time.

Despite being *informed* at the same time, users were in a different phase of their membership “age” on the website (i.e., number of weeks since initial joining). Site activity varies substantially across membership ages: for most users, the number of activities increases over the first few weeks, and then decreases, with varying slope contours; other users increase their number of activities over time, and others might have distinct patterns of activities throughout their membership lifetime. The focal point is that the breach itself occurred at different points in these trajectories; although the shock hit all users, it did so at different points in their experience and

consequent activity pattern on the site. While users differ in their trajectories, having a relatively large number of users in each cohort (average number of users per cohort is 2,161; standard deviation 295; smallest cohort has 1832 users) permits matching users’ trajectories on the website, across the different cohorts. This in turn allows us to construct “Temporal Causal Inference”: for each cohort, we compare their behavior to a group of users who joined in previous weeks. For this group of users who joined earlier, which will be referred to as the “control group”, we observe a larger number of weeks of activity prior to the announcement of the breach. The cohort who joined later, referred to as the “treated group” (the treatment being the announcement of the breach), was affected by the announcement of the data breach earlier in their lifetime on the website, compared to the control group. In particular, let  $J_T$  denote the cohort of users who joined  $T$  weeks **before** the data breach was announced. For this cohort, we observe  $T + 3$  weeks, where the last 3 weeks are after the breach announcement. Let groups  $J_1, \dots, J_{T-3}$  be all the cohorts that joined at least 3 weeks before  $J_T$ . From these cohorts, we use the first  $T + 3$  weeks of activity since joining the website and form a control group  $J_T^C$ .<sup>10</sup>

It is important to note that all users in the control group were also affected by the breach. However, they were only affected later in their lifetime (membership age) on the website. For this control group,  $J_T^C$ , the time of (not receiving the) treatment will still be  $T$ , and the 3 weeks to follow (which were all before the breach) will aid us in predicting *what would have been* the (expected) behavior of the treatment group, had the breach not been announced.

We illustrate the construction of the groups in Figure 2.2: the upper panel illustrates the average number of activities for one treatment cohort, and three earlier cohorts, over time; the earlier cohorts are formed into a single control group. The lower panel illustrates that, if the X-axis variable is Membership Age (i.e., not cal-

---

<sup>10</sup>In practice, in order to have similar number of users, we take as control group the last 5 cohorts that joined prior to the treated group. Results are robust to any number of cohorts larger than 3.

endar time), the groups are comparable, except that the control group was not yet exposed to the breach. For the control group, we use only the data *up to the treatment*.

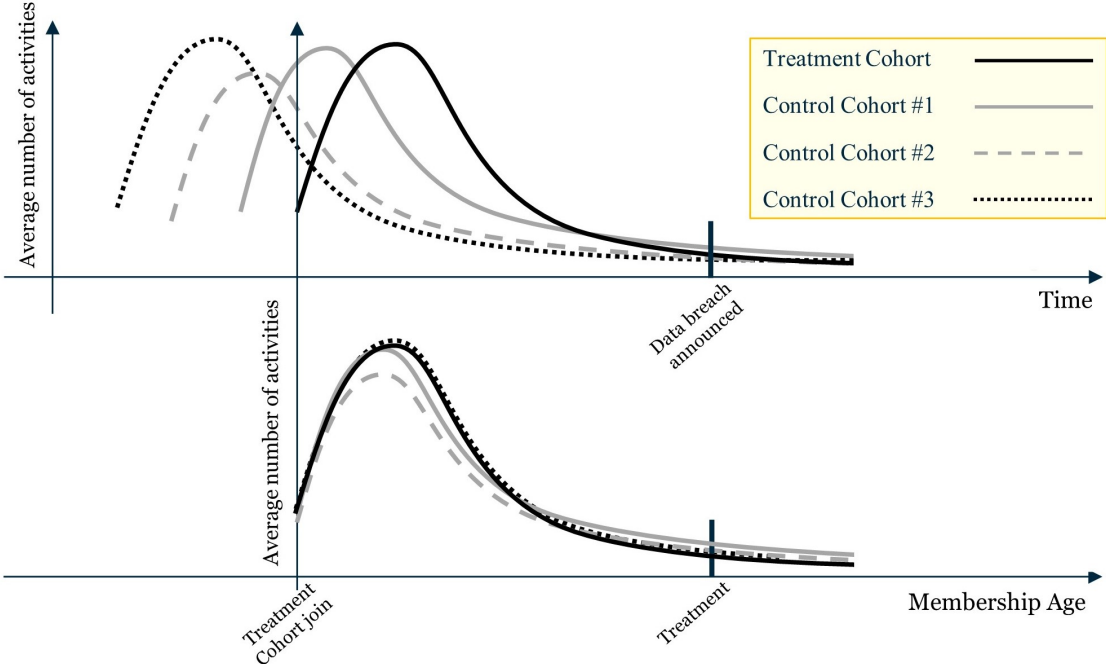


Figure 2.2: Illustration of Temporal Causal Inference, with one treatment cohort, and 3 control cohorts.

We repeat the construction of the control and treatment groups 21 times: for each of the cohorts who joined in a specific week 1-6 months prior to the data breach, we employ all the users who joined at 3-8 weeks prior to them. Such repetition results in 21 different control groups – one for each of the 21 treatment groups. This explicitly means that almost all cohorts serve multiple times as part of a control group (for varying lengths of their membership age), and that all cohorts except for the first three (to which we have no one in our data sample to serve as a control group), serve only as a treated group.

Figure 2.3 shows a visual comparison between control and treatment groups, as constructed by Temporal Causal Inference (TCI), while Figure 4 depicts only the last three weeks prior to the announcement. Figure 2.5 shows the average number of activities for each of the three weeks prior and post announcement, across all

cohorts. This figure provides “model-free evidence” demonstrating the change in the probability of being active in each of the activities on the website, in the weeks following the announcement.

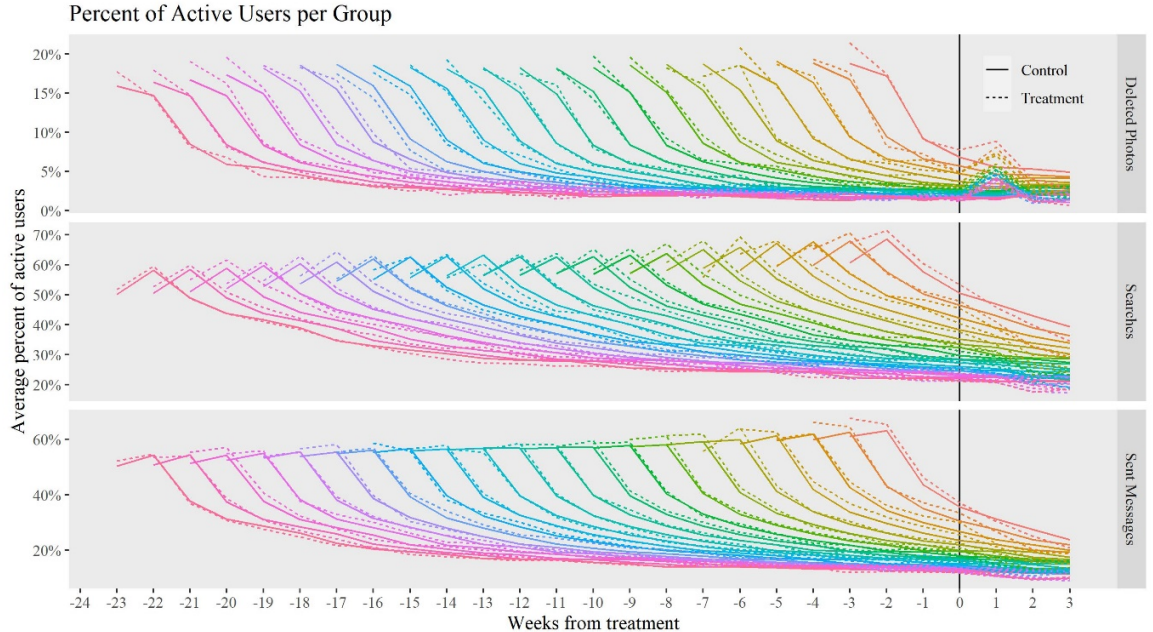


Figure 2.3: Percent of active users per Control/Treatment group, as a function of week from treatment. Solid vertical lines indicate the announcement of the breach. Different colors indicate week of joining the website for each of the 21 treatment groups.

### 2.4.2 Causal Inference Assumptions With TCI

Our causal inference mechanism builds on the “potential outcome framework”, a term attributed to Neyman (Rubin, 2005). According to the potential outcome framework, any causal inference problem relies on two quantities:  $Y_i^{(0)}$  and  $Y_i^{(1)}$ , the outcome of unit  $i$  without and with receiving the treatment  $W_i \in \{0, 1\}$ , respectively. The typical measure of the treatment effect is  $Y_i^{(1)} - Y_i^{(0)}$ . A fundamental problem of causal inference is that, in any experiment – may it involve randomly assignment or not – the researcher cannot ever observe both  $Y_i^{(0)}$  and  $Y_i^{(1)}$ , as the user is either exposed to the treatment, or not. Therefore, the researcher should estimate the



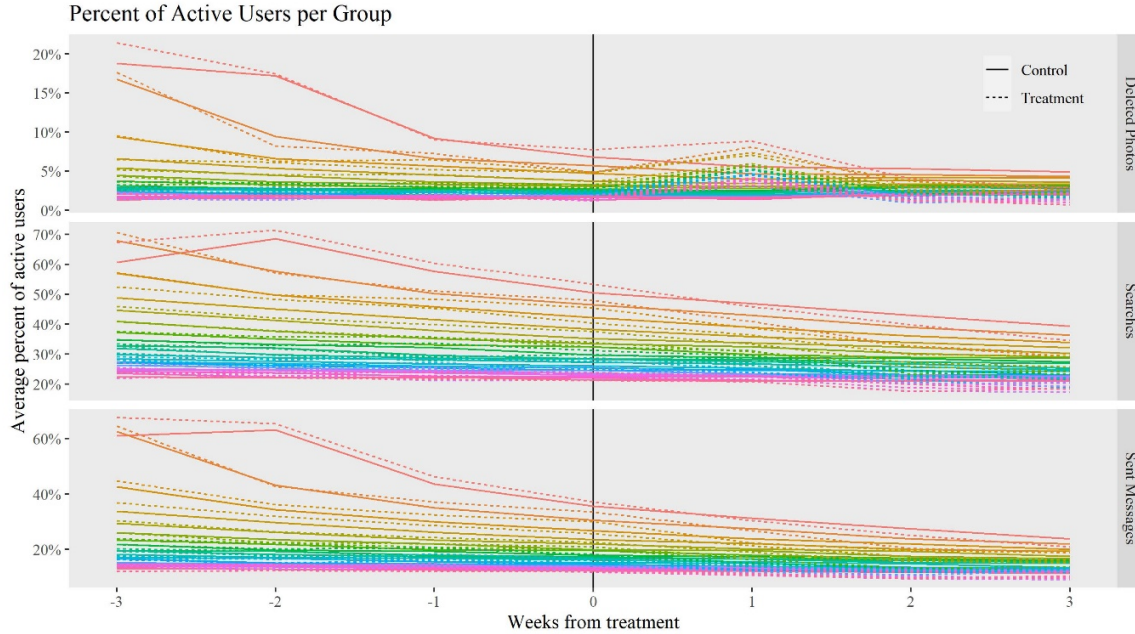


Figure 2.4: Percent of active users before and after the announcement. Solid vertical lines indicate the announcement of the breach. Different colors indicate week of joining the website for each of the 21 treatment groups.

missing quantity – referred to as the potential/counterfactual outcome – either  $\hat{Y}_i^{(0)}$  or  $\hat{Y}_i^{(1)}$ , and use the counterfactual outcome to infer the causal effect. In order to do that, we should also estimate the probability to be treated, which we denote  $\hat{w}_i \in (0, 1)$  (the restriction of not being 0 or 1 is explained in Section 2.4.2.3).

In the case of a data breach that affected – even if merely by attention – the entire website population, there is no random assignment to treatment, and we therefore construct TCI. In this section, we illustrate how TCI can assist in adhering to Causal Inference assumptions, in order to be able to measure the treatment effect of any exogenous shock. In our case – the measurement of the treatment effect of the announcement of a data breach. We will briefly review four main assumptions of Causal Inference: Stable Unit Treatment Value Assumption (SUTVA), Conditional Independence Assumption, Overlap Assumption and Exogeneity of Covariates Assumptions. For each, we will show how TCI helps in identifying the treatment effect and assures adherence to the assumption, in the case of an exogenous shock. In cases where there

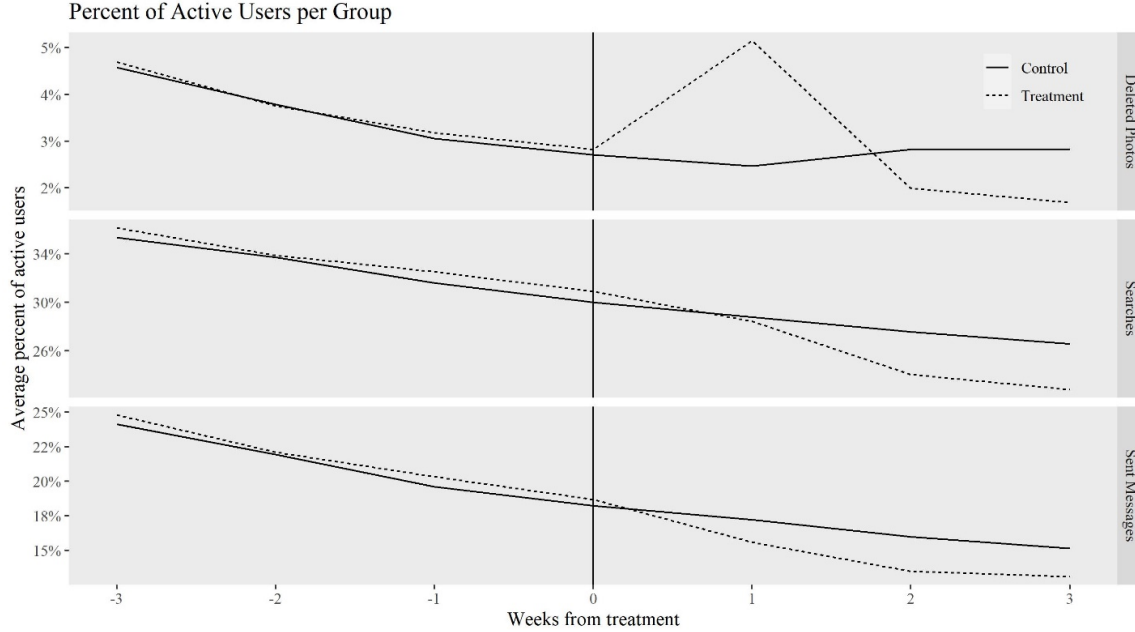


Figure 2.5: Average percent of active users across all cohorts.

might be additional concern for clear identification of the treatment effect, we will discuss possible solutions we propose to overcome them.

#### 2.4.2.1 SUTVA (Stable Unit Treatment Value Assumption)

The SUTVA assumption (Rubin, 1980) comprises two conditions:

**a. No interference between units (Cox, 1958).** Neither  $Y_i^{(1)}$  nor  $Y_i^{(0)}$  is affected by the treatment assignment any other unit received:  $Y_i^{(W_i)} \perp Y_j^{(W_j)}$  for any two users  $ij \in 1, \dots, N$ . In the case of measuring the effects of an exogenous shock, this assumption means that if one user/customer is exposed to the shock, this should not affect the outcome of any other user. By design of TCI, assignment to treatment or control groups is based on the time of joining the website. Since the outcomes do not occur at the same time, no two users who were in the *different* groups – control and treatment – are affected by the other’s treatment or lack thereof. That is, for every set  $\{i, j\}$   $Y_i^{(1)} \perp Y_j^{(0)}$ . On the other hand, behavior following the data breach of one treated user, might affect another user, that is, it is possible that  $Y_i^{(1)} \perp Y_j^{(1)}$  for

some *treated* users  $ij$ . This is because of the nature of the website: a match-making website, where one user’s change in behavior (e.g., stopping using the website) might affect another user’s behavior (e.g., enjoys the website more, now that there is less “competition” on the website). In addition, due to network effects, it might be the case that  $Y_i^{(0)} \perp Y_j^{(0)}$ . Since our objective is to measure the effect of the data breach, and in general – the effect of other exogenous shocks – this possible dependency is not a concern; rather, it **should** be part of the estimation of the treatment effect.

**b. No hidden versions of treatments.** Also known as the Consistency Assumption. This assumption states that  $Y_i = Y_i^{(1)} \cdot W_i + Y_i^{(0)} \cdot (1 - W_i)$ , the outcome that would be observed for treated unit would be  $Y_i^{(1)}$  and for control unit  $Y_i^{(0)}$ ; i.e., nothing, except for the treatment, affects the outcome. This assumption usually cannot be tested. As in almost every causal inference scenario other than perfectly randomized control trials (only those which are repeated in multiple occasions. places, and with large enough sample), there might be other, unobserved events that affect users’ behavior. For example, during the time period of the breach there might have also been a change to the platform or a holiday that otherwise affected users’ outcome  $Y_i^{(1)}$  independently of the data breach. To the best of our knowledge, and according to data from prior periods, at the time of the breach there was no change to the platform, no notable holiday, no other such events, except for the data breach itself. As for users in the control groups, the construction of TCI – to include multiple cohorts as “control cohorts”, each joining at a different time – mitigates the likelihood of having any unrelated event affecting  $Y_i^{(0)}$ . This is because, by using multiple control cohorts, and by repeating TCI for multiple treatment cohorts, such events are smoothed through the average of all other control cohorts. Moreover, we show subsequently, in a series of analyses, that, once constructing TCI, the control and treatment groups are indistinguishable in their behavior prior to the treatment. Nevertheless, in order to estimate the individual treatment effect, we complement our Temporal

Causal Inference with Causal Forests methodology, thus matching individual users to an ensemble of users from the control group.

As noted by Rubin (2005), under randomized controlled trials (RCT), there is no need for any assumptions other than SUTVA. However, we cannot avail of an RCT, and therefore we need to assure our models have properties that are “given” in RCTs. We therefore proceed with showing that TCI allows for recovery of the treatment effect of an exogenous shock, while holding the necessary following assumptions:

#### 2.4.2.2 Conditional Independence (or Ignorability) Assumption

Assignment to Control/Treatment Groups are random, conditional on  $X$ :

$$\Pr(W_i | X_i, Y^{(0)}, Y^{(1)}) = \Pr(W_i | X_i, Y_{obs})$$

where  $Y_{obs}$  is the observed outcome. This assumption was later extended to “unconfoundedness” assumption (Rubin, 1990), which entails that there is no need to control for  $Y_{obs}$ :

$$\Pr(W_i | X_i, Y^{(0)}, Y^{(1)}) = \Pr(W_i | X_i)$$

The data breach, as an exogenous shock, affected all users. However, because some users joined the website earlier, they were affected by the breach at a later point in their membership age, for unobserved reasons that might be confounded with the treatment effect (e.g., if for some reason, people that are more active, selected to join in a specific month). Therefore, the construction of TCI might not overcome this. In order to assure that the control group has similar behavior to that of the control group, prior to the treatment, we test the parallel trend assumption using a Granger Test (Granger, 1980)<sup>11</sup>. We find that, across all treatment groups and across

---

<sup>11</sup>Granger Causality Test is a common way to compare two time series. Despite “causality” in its title, it is a well-known test of predictability, or “temporal relatedness”, of one time series to another and should not be misconstrued as a test of causality. We implemented a bi-directional test of predictability of the control group on the treatment group. We found that for all treatment

all activities, the control groups we constructed using Temporal Causal Inference are acting as perfect predictors for the behavior of their respective treatment group, and vice versa. We also conducted a Kolmogorov-Smirnov Test (Massey Jr, 1951) to verify that the control and treatment groups do not differ in distribution. Specifically, for each activity and for each treatment-control pair, we compute a cumulative sum of the average percent of active users prior to the treatment and divide it by the groups' maximum cumulative sum; this effectively creates a CDF-like timeline, as shown in Figure 2.6. We then conduct a Kolmogorov-Smirnov Test to test whether the CDF-like timelines differ between the control and treatment groups. We find that, throughout all activities and treatment groups, the trends are indistinguishable from those of the respective control groups.

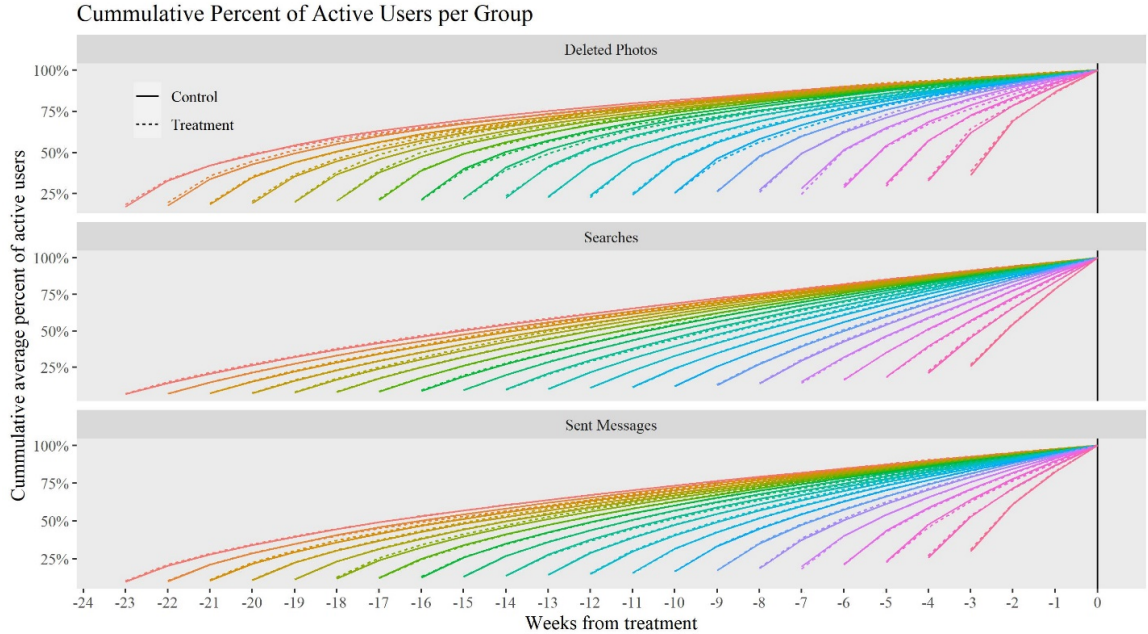


Figure 2.6: Comparison of control and treatment groups: CDF-like comparison of treatment and control groups. Each line is the cumulative sum of the percent of active users for this group, divided by the maximum cumulative sum for this group (therefore always gets to 100%, and starting “higher” for cohorts with shorter timelines, due to this).

---

groups and for all activities, the time series was statistically indistinguishable (the null hypothesis of no-prediction was rejected with  $p < 0.001$  across all tests).

Even though we established that there are parallel trends between the groups, in order to estimate individual treatment effects while also acknowledging that users might have varying timelines, we will add a second step to our causal inference method, one that will account for possible differences in in the timelines of users prior to the treatment. In short, we will match users in the control and treatment groups, nonparametrically, using Temporal Causal Forests.

### 2.4.2.3 Overlap Assumption (or “Common Support” Assumption)

According to this assumption, the propensities to be treated are strictly between 0 and 1:

$$0 < \Pr(W_i = 1 | X_i = x) \equiv \hat{w}(x) < 1$$

In TCI, due to the nature of the exogenous shock, every user had some propensity to have been treated in any week during his period on the site; evidently, all users were treated. Nevertheless, in the “Causal Forests” section, we will further test this assumption by observing the estimated  $\tilde{w}(x)$  – the propensity to be treated – estimated by Local Linear Forests.

### 2.4.2.4 Exogeneity of Covariates Assumption

This assumption states that the covariates are not affected by the treatment:

$$X_i^{(1)} = X_i^{(0)}$$

The data breach was an exogenous shock that, to the best of our knowledge, was not explicitly predicted by any user or employee. Even if there are users that did expect something of this sort to happen, this is likely to be true in both the control and treatment group, and therefore should be overcome using TCI. Therefore, it is not of a concern in identifying the treatment effect.

To summarize, Temporal Causal Inference aided us in finding proper control groups to measure the treatment effect on the exogenous shock of the data breach. We found that the construction of these groups ensures all necessary and sufficient causal inference assumptions hold. In the next section, we will introduce Causal Forests – a non-parametric method that will further allow us to find a matching control for each user in our treated groups. Temporal Causal Forests will also allow us to measure the heterogeneity in the treatment effect.

## 2.5 Temporal Causal Forests

Toward the aim of determining the heterogeneous effects of the announcement of the breach, we take the control and treatment groups created via TCI and estimate individual treatment effects using a nonparametric, forest-based method – Temporal Causal Forests. We also ran several semi-parametric and parametric models. Simulation studies were conducted to assure that the method chosen was able to recover simulated treatment effects. All methods will be described in the robustness-checks and simulation sections.

### 2.5.1 Comparing Control and Treatment Users

Since our goal is to estimate heterogeneous treatment effects, we must carefully construct the comparison based on individual trajectories (probability to be active in each week of membership on the website). We have 8-24 weeks of observations for users in various treatment groups, based on their times of joining the website. We use this timeline and compare it to that of similar users in the control group, in order to estimate what would have been the probability to be active, following the treatment (announcement of the data breach), if it were to not occur. In order to do so, we use a forest-based method, specifically, Temporal Causal Forests. It is important to note that Temporal Causal Inference and Temporal Causal *Forests* comprise two

separate, sequential procedures: while TCI generates two groups (control and treatment), Temporal Causal Forests, denoted TCF, estimate the heterogeneous treatment effect, by estimating *what would have been the number of activities had the user been in the opposite group*. Simply put, the difference between the estimated number of activities, and the observed one, provides an estimate of the effect of the breach. The second step, TCF, can be seen as a nonparametric propensity-score matching mechanism; predictions are made nonparametrically based on the entire corpus of data, so that each user in the treatment group will be fit with the suitable counterpart in the control group.

We adapt the Causal Forest (CF) (Wager and Athey, 2018) model using Generalized Random Forests (GRF) implementation (Athey and Wager, 2019). In recent extensive simulation study, Causal Forests in this implementation have been found to show strong performance, under all tested settings (Knaus et al., 2021).

We introduce two changes to the original Causal Forest method. These changes are both internal and external to the estimation of the treatment effect, and were found, in a series of simulation studies we ran (to be described in Section 2.7.5), to give the best results in terms of RMSE and ability to recover heterogeneous treatment effects, both in synthetic data and on our dataset. These changes are:

1. In most cases, the use of the Causal Forest is to generate groups that are equivalent in their propensity to be treated, and to compare between the users within each group. In our case, however, we choose the parameters  $X_i$  to include both psycho-demographic covariates, and, more importantly, the *time trend*, as will be explained below. This leverages the traditional Causal Forests framework to group users based on their pattern of activities throughout time, resulting in groups within the control and treatment groups that are relatively homogeneous in respect to their *time trend*. In other words, TCF allows us to assess individual treatment effects by estimating a counterfactual time trend



of their activities. Therefore, TCF allows treatment and control groups to be compared while verifying that the users have similar time trends before the breach announcement (i.e., the treatment).

2. In order to improve Causal Forests, we estimate its “nuisance parameters” (to be explicated later) using Local Linear correction, via Local Linear Forests (Friedberg et al., 2020).

### 2.5.2 Construction and Estimation – Temporal Causal Forests

The TCF methodology, as carried out here, consists of sequential application of four nonparametric forest-based methods: Random Forests, Causal Forests, Generalized Random Forests and Local Linear Forests, each building atop its predecessors. We now briefly explain each of these components, omitting widely known details from the core Random Forests literature (e.g., Breiman (2001)):

**a. Random Forests** is a supervised machine learning method aimed at estimating a prediction  $\hat{\mu}(x)$  for a vector of covariates  $X_i = x$ . The estimation can be seen as an “ensemble method”, by taking the average of all regression/decision trees. Each decision tree  $b$  is constructed so that the leaves  $L_b$  will include observations that have similar set of covariates. Specifically, we present here “Honest Forests”: for each tree  $b \in \{1, \dots, B\}$ , where  $B$  is the number of trees in the forest, draw a subsample  $S_b$ , referred to as the training sample, in the size of half of the population (the size can be tuned). Grow the regression tree by recursively splitting so that the error function (to be defined for each problem separately) will be optimized (usually minimized). After training the forest for each user with set of covariates  $x$  not in  $S_b$ , make out-of-bag predictions on the response variable,  $\hat{\mu}(x)$ :

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^N Y_i \frac{\mathbb{I}\{X_i \in L_b(x), i \notin S_b\}}{|\{i : X_i \in L_b(x), i \notin S_b\}|} \quad (1)$$

where  $L_b(x)$  is the leaf of the  $b$ -th tree, to which the set of covariates  $x$  correspond, according to the splitting rule. Wager and Athey (2018) showed that, when using Random Forest with “honesty” – that is, by using B trees, where the training set is randomly chosen for each – one can derive the asymptotic distribution of the response variables, thus allowing us to get both mean and variance of individual estimates. In the sequel, we assume the “honesty” property, and remove notation of  $S_b$  for simplicity.

**b. Generalized Random Forests (GRF).** Whereas Random Forests can be seen as an ensemble method – average of predictions made by individual trees – Athey et al. (2019) propose that it can be seen as an adaptive kernel method, in a Generalized Random Forest:

$$\begin{aligned}\hat{\mu}_{grf}(x) &= \sum_{i=1}^N \alpha_i(x) \cdot Y_i \\ \alpha_i(x) &= \frac{1}{B} \sum_{b=1}^B \alpha_{bi}(x) \\ \alpha_{bi}(x) &= \frac{\mathbb{I}\{X_i \in L_b(x)\}}{|L_b(x)|}\end{aligned}$$

Therefore, the weights  $\alpha_i(x)$  are higher for the observations that appear more often in the same leaf as  $x$  – and are thus “closer to it” – whatever the measure of splitting mandates to be as “closer” (hence the term “generalized” in the name of the method). Note that by construction,  $\sum_{i=1}^N \alpha_i(x) = 1$ , and  $\alpha_i(x) \geq 0$ .

**c. Causal Forests (implemented using GRF).** Here we will follow the GRF method for estimating Causal Forests, with honesty. Assume  $(X_i, W_i, Y_i)$ , where  $X_i$  is defined as earlier,  $W_i \in \{0, 1\}$  is the treatment assignment of user  $i$ , and  $Y_i = Y_i(W_i)$  is the observed outcome. For each user we observe either  $Y_i(1)$  or  $Y_i(0)$ , but not both,

and we aim to estimate the treatment effect, which can be simplified into:

$$\hat{\tau}(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$$

Until now, we have not defined the splitting rule by which the tree splits were constructed. Usually, it is set to minimize the Sum-of-Squared Error between the observed and predicted outcome: Consider a parent node  $P$  with  $n_P$  observations,  $(X_{i1}, Y_{i1}), \dots, (X_{in_P}, Y_{in_P})$ . For each candidate pair of child nodes,  $\{C_1, C_2\}$ , let  $\bar{Y}_1, \bar{Y}_2$  be the corresponding mean of  $Y$  in that leaf. The chosen pair of child nodes will be those that minimize:

$$\sum_{i: X_i \in C_1} (Y_i - \bar{Y}_1)^2 + \sum_{i: X_i \in C_2} (Y_i - \bar{Y}_2)^2$$

However, in causal inference, we do not observe both  $Y_i(1)$  and  $Y_i(0)$ , but rather only one of them. Therefore, we cannot compare the predicted outcome to the quantity we do not observe. In order to solve this, in Causal Forests the splitting rule is based on pseudo-outcomes: within each leaf, the splitting rule is constructed so that the *propensity to be treated* for those in the control and those in the treatment group, conditioning on the covariates, is similar. In particular, let the pseudo-outcome be:

$$\rho_i = \frac{\left( (W_i - \bar{W}_P) \left( Y_i - \bar{Y}_P - \hat{\beta}_P (W_i - \bar{W}_P) \right) \right)}{\text{Var}_P(W_i)}$$

where  $\bar{W}_P, \bar{Y}_P$  are the averages taken over the parent node  $P$ , and  $\hat{\beta}_P$  is the least-squares regression solution of  $Y_i$  on  $W_i$  in the parent node  $P$ .  $\text{Var}_P(W_i)$  is the variance of the treatment in the parent node:

$$\text{Var}_P(W_i) = \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i: X_i \in P\}} \left( (W_i - \bar{W}_P) \otimes (W_i - \bar{W}_P)^T \right)$$

The splitting rule is then calculated along the gradient of the mean difference with the pseudo-outcomes, to maximize homogeneity in the propensity to be treated in each leaf. Specifically, the parent node will be split to two leaves  $\{C_1, C_2\}$  that maximize:

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^2 \frac{1}{|\{i : X_i \in C_j\}|} \left( \sum_{\{i: X_i \in C_j\}} \rho_i \right)^2$$

This splitting criterion is used recursively, therefore generating a tree, and later a forest. The weighing mechanism specified in the description of GRF then takes place.

In the notion of “Causal Inference”, all the above decisions merely mean that the splitting is done so that there will be similar propensity to be treated for each observation within each leaf (Guo et al., 2021). This assures that despite having possible differences between the control and treatment groups, these are being minimized in each leaf – in order to have a homogeneous control and treatment group in each leaf.

**Nuisance Parameters in GRF Causal Forests.** In order to improve efficiency and be more robust to confoundedness, Athey et al. (2019) show that it is possible to maintain accuracy and asymptotic inference by first regressing out (locally centering) the effects of  $X_i$  on the outcomes that are used to perform the optimization. In order to do so, they introduce nuisance parameters:

$\hat{w}(x) = \mathbb{P}[W_i | X_i = x]$  is the propensity to be treated, and

$\hat{y}(x) = \mathbb{E}[Y_i | X_i = x]$  is the expected outcome, marginalizing over the treatment.

Then, center the outcome and treatment  $\tilde{Y}_i = Y_i - \hat{y}^{(-i)}(X_i)$  and  $\tilde{W}_i = W_i - \hat{w}^{(-i)}(X_i)$ , where the  $(-i)$  superscript denote out-of-bag estimates of  $\hat{y}$  and  $\hat{w}$ , computed without using the  $i$ th observation, as explained in the definition of “honest forests” above.

Using these quantities, the forests can be run on the pair  $\tilde{Y}_i \tilde{W}_i$ . Then, after

constructing the forest, we can estimate the treatment effect via

$$\hat{\tau}(x) = \frac{\sum_{i=1}^n \alpha_i(x) \cdot \tilde{Y}_i \cdot \tilde{W}_i}{\sum_{i=1}^n \alpha_i(x) \cdot \tilde{W}_i^2}$$

As stated earlier, Wager and Athey (2018) derive asymptotics of  $\hat{\tau}(x)$ , and thereby estimated individual treatment effects’ mean and variance,  $\tilde{\sigma}_\tau(x)$ . The variance is estimated using the infinitesimal Jackknife method, also known as the non-parametric delta model (Wager et al., 2014). We refer the reader to section 6 of Athey et al. (2019) and Athey and Wager (2019) for further detail on the construction of the Causal Forest in Generalized Random Forests settings.

**d. Local Linear Forests.** The last step of our forests collection will be used in the estimation of the nuisance parameters,  $\hat{w}(x)$  and  $\hat{y}(x)$ . Whereas in Athey et al. (2019) these quantities were estimated via regression forests, Friedberg et al. (2020) demonstrates how to achieve better accuracy by estimating  $\hat{w}(x)$  and  $\hat{y}(x)$  using Local Linear Forests. Local Linear Forests build on Generalized Random Forests, and add a layer of linear regression to exploit smoothness of the outcome, and to correct for potential misalignment between a test point and its neighborhood. It may also aid in resolving instances of unbalanced data and noise.

### 2.5.3 Temporal Causal Forests

Until now, we have described the various forest types used in our estimation. We now explain the step-by-step process of both stages of our analysis. We create a Temporal Causal Forest for each Treatment Group, along with its respective Control Group, for each week after the treatment  $p \in \{1, 2, 3\}$ , and for each type of activity (Sent Messages, Searches, Deleted Photos). The treatment effects of each user in the Treatment Groups are estimated once for each user in the treatment groups, when

this user “acts” as a treated user.

The TCF algorithm illustrates the steps (in pseudo-code) taken in each Temporal Causal Forest:

```
X = is(x_it > 0)
Y = is(y_ip > 0)
W = is(i in treatment group)
Y.hat = local_linear_forest(X, Y)
W.hat = local_linear_forest(X, W)
tau.hat.stage.1 = causal_forest(X, Y, W, Y.hat, W.hat)
```

In each Temporal Causal Forest:

1. Estimate nuisance parameters,  $\hat{y}(x)$  and  $\hat{w}(x)$  for all users, using Local Linear Forests. We tune the parameters for both local linear forests using the built-in tuning, which bootstrap over the available parameter space and optimally finds the suitable scaling parameters. We construct 2K honest trees<sup>12</sup>, and make predictions using Honest Trees as described above.
2. Given  $\hat{y}(x)$  and  $\hat{w}(x)$  from 1, build a Causal Tree on GRF, which classifies users from the Control and Treatment groups, based on their set of features,  $x = X_i$ . As noted earlier, in the case of Temporal Causal Forests developed here, the set of covariates (features) used are the timeline of users (an indicator: whether the user engaged in this activity each week) before the treatment ( $X_{it}$ ), along with such psycho-demographic features as age, marital status, and privacy preference. The parameters used in creation of the trees are estimated using the built-in tuning, so that the optimization of the parameters is carried out by bootstrapping as described above. We use 2K honest trees.

---

<sup>12</sup>This is the recommended number of trees for this size dataset. Results were robust to other specifications of relatively large number of trees.

3. The results from running the TCF is an individual estimate of treatment effects and variances around them. Therefore, each user in a control group has, for 3 types of activities, 3 weeks after the announcement, a measure of the change in probability of being active, along with variances of these estimates.

**Population Mean and Variance.** To estimate both the mean treatment effect across the population and the associated variance, we proceed as follows: let  $(\hat{\tau}_i, \sigma_{\hat{\tau}_i}^2)$  be the individual estimated treatment effect and variance, respectively, as computed by Temporal Causal Forest (removing the  $t$  subscript for conciseness). The mean treatment effect is then the average across all users who were treated:

$$\bar{\tau}_{treated} = \frac{\sum_{i \in treated} \hat{\tau}_i}{N_{treated}}$$

The variance around this estimation is constructed from both the uncertainty (variance) around each individual estimate ( $\sigma_{\hat{\tau}_i}^2$ ), and the uncertainty around the mean of all individual estimates:

$$Var(\bar{\tau}) = \frac{\sum_{i \in treated} [\sigma_{\hat{\tau}_i}^2 + (\hat{\tau}_i - \bar{\tau}_{treated})^2]}{N_{treated}^2}$$

#### 2.5.4 Sources of Heterogeneity in the Effect of the Information Shock

As indicated in the literature review, users may have varying reactions to data breaches, owing in part to different expectations regarding website’s duty to protect their personal information. In addition, in settings like that presented here, the data announced as breached could well reveal an active search for an extramarital affair. Although results stemming from this specific dataset may be of interest in itself – due to the nature of the data breach and for the investigation of reactions to severe privacy invasion, TCF may be used in other settings to afford clearer understanding of possible reasons that make individuals more, or less, reactive to a variety of exogenous

shocks.

A specific example in our data setting concerns whether married people would be differentially affected by the breach, since they have, *ceteris paribus*, more to lose compared with single ones (as a reminder, using a survey we ran on the entire population, we verified that there is no clear bias in disclosing marital status on the website). In such cases, possible sources of heterogeneity may be explicable by users’ willingness to be more “public” in their profile, assessed by the available indicator of whether the user had a public photo on the website<sup>13</sup>. Those with public photo run the risk of revealing their identity on the website even before the breach (e.g., depending on how identifiable they are in the photo), potentially because they entrusted the website’s security protocols.

In addition to these, in our setting, it might be possible that users in different treatment groups that were affected by the data breach at different membership ages will be affected by the data breach announcement in distinct ways. To locate observed sources of heterogeneity in the response to the breach announcement, we complement Temporal Causal Forest results with a linear regression. Specifically, for user  $i$  with treatment effect tuple,  $\hat{\tau}_i$  (i.e., for each  $t \in \{post1, post2, post3\} = \{1, 2, 3\}$ , after the breach was announced), we regress the effect of the announcement for that week, as follows:

$$\hat{\tau}_{it}(X_i) = \beta_0 + \beta_C Cohort_i + \beta_M Married_i + \beta_P Private_i + \epsilon_{it}$$

For ease and consistency of interpretation, all independent variables – cohort, marital status (married / single), and an indicator for whether the user had a public photo on his profile – are mean-centered and standardized, e.g., the intercept refers to the “centroid case” within the data.

---

<sup>13</sup>The photo is public only to registered users, but it is not necessary to pay in order to register. Therefore, theoretically any person who was interested in seeing photos of users on the website could have joined, and seen photos of users willing to share them.



## 2.6 Results

We first discuss the average treatment effect, and then possible sources of heterogeneity.

### 2.6.1 Average Treatment Effect

The average effects of the announcement of the data breach on deleted photos, which are depicted in Figure 2.6 and Table 2.1, suggest a substantial increase: 2.8% more of the users (which is double the baseline) deleted at least one photo immediately after the announcement (Week 1). This seems reasonable, considering that users tried to “cover their traces”, after realizing that their personal data were no longer secure on the site. While hardly a remedy for their data security loss, it was at least a protective action within their own control (one we can observe), to lower the probability of further identification. In the following weeks, there is a significant decrease in the probability of deleting photos, compared to before the announcement; this suggests that whoever wanted to delete their photos did so immediately after the announcement, and therefore had fewer photos to delete in the following weeks.

Table 2.1: Temporal Causal Forests Mean and Standard Deviation

		Week 1	Week 2	Week 3
Sent Messages	Mean	-0.016***	-0.027***	-0.022***
	SD	(0.001)	(0.001)	(0.001)
Searches	Mean	-0.006***	-0.039***	-0.037***
	SD	(0.001)	(0.001)	(0.001)
Deleted Photos	Mean	0.028***	-0.003***	-0.005***
	SD	(0.001)	(0.001)	(0.001)
# Observations		49,993		
Note: *p <0.05; **p <0.01; ***p <0.001				

In contrast to deletion of photos, for searches and sent messages, on average users decrease both of these activities, for all three weeks observed after the announcement. For searches, it seems that the effect is mainly manifested in the second and third

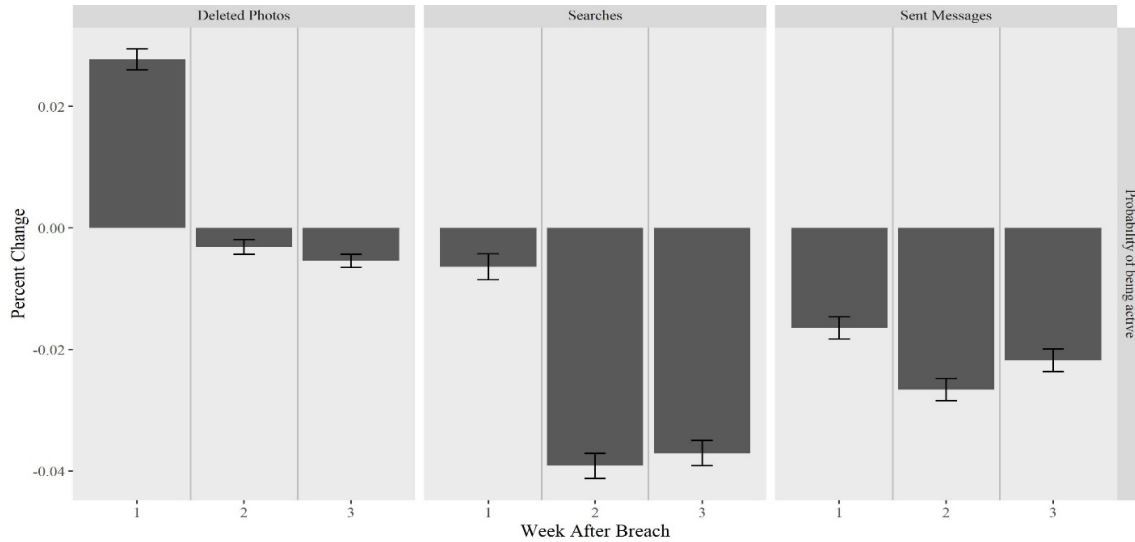


Figure 2.7: Average treatment effects and 95% confidence interval of the change in probability of being active.

week after the announcement, in terms of decrease in probability of being active (3.9% and 3.7% decrease, respectively).

For sent messages, results suggest some attenuation (i.e., “less decrease”) in Week 3, in terms of probability of being active, compared to the week before (2.7% in week 2, vs 2.2% in week 3). This suggests the effect is waning over time, in line with Rosati et al. (2017), who found a short-term reduction in bid-market price following a data breach, but no long-duration effect; our three-week post-breach data window do not allow this to be verified.

In order to assess the model fit, Table 2.1 presents the estimated counterfactual, along with the model free evidence of TCI. The results suggest that TCF correctly estimates the percent change in probability of being active.

Whilst in Figure 2.8 we presented the mean treatment effects, taking into consideration the individual-level error and population variance, we note that there are varying reactions to the data breach. The distributions of the individual treatment effects (taking into consideration only the mean individual estimates) are displayed in Figure 2.9. This illustrates the range of estimated reactions to the data breach, where

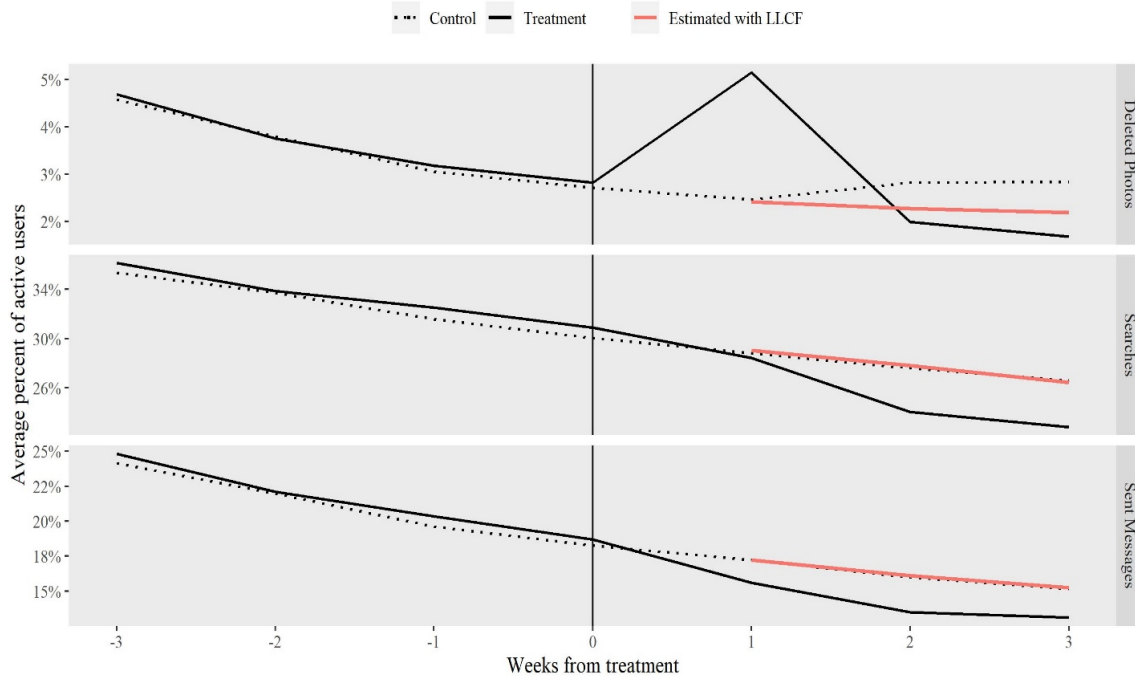


Figure 2.8: Percent of active users per Control/Treatment group with estimates, as resulted by Temporal Causal Inference, for the treated group (solid line) and control group (dashed). In red: the estimated counterfactual percent of active users as estimated by Temporal Causal Forests. The treatment effect is the difference between the observed and counterfactual behavior of the treated group.

some users increased the likelihood of being active, others have decreased it. Understanding the reasons for the various reactions may assist policymakers and businesses to tailor their messages of post-breach protective measures to populations that may be less likely to react. Businesses who are affected by the data breach may be able to assist their customers in protecting themselves, as well as to understand where harm to trust was the most extreme. In the next section, we will uncover some of the sources for this heterogeneity in reactions.

### 2.6.2 Observed Sources of Heterogeneity

We now aim to identify sources of heterogeneity in the treatment effect. To do so as transparently as possible, we regress the individual treatment effects,  $\hat{\tau}_{it}(X_i)$ , on the key input covariates, *Cohort*, *Married*, and *Private* (no public photo on the

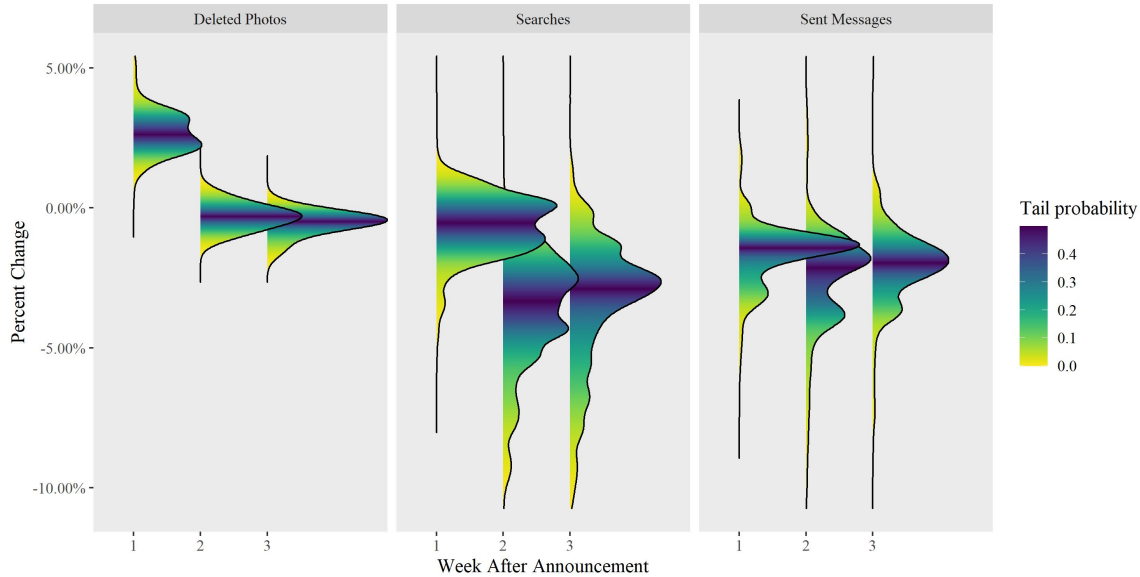


Figure 2.9: Distributions of the heterogeneity in treatment effects – mean individual changes in the probability of being active in each activity on the website. Each panel is for a different activity, each column is for different week following the announcement of the data breach, and colors indicate the tail probability.

website). Note that this is fully analogous to “Level II” of a hierarchical linear model, except the dependent variables here are the outputs of the TCF, which we seek to explain. Because all covariates are mean-centered and standardized, the intercept corresponds to the average estimated treatment effect, as captured by TCF, and is similar to the population estimates presented above<sup>14</sup> in Figure 2.7. The other “ $\beta$ ” coefficients capture marginal effects of *deviations* from the average effect of the data breach, for those who are, respectively, “newer” on the website (joined later), married and private (had no public photo). The coefficients of searches and messages, presented in Figure 2.10, (along with the added common intercept that is presented in Figure 2.11 for ease of interpretation with the average treatment effect) indicate that newer users (higher “cohort”) were less active – that is, reduced their activities more than users that were on the website for longer duration, in all weeks and in all

<sup>14</sup>In lieu of tabular results, which are numerous and available in Appendix 2.9.1, we present analyses for outcomes of interest – i.e., the effects of the breach and covariates associated with them – visually via mean effects and associated (95%) confidence intervals.

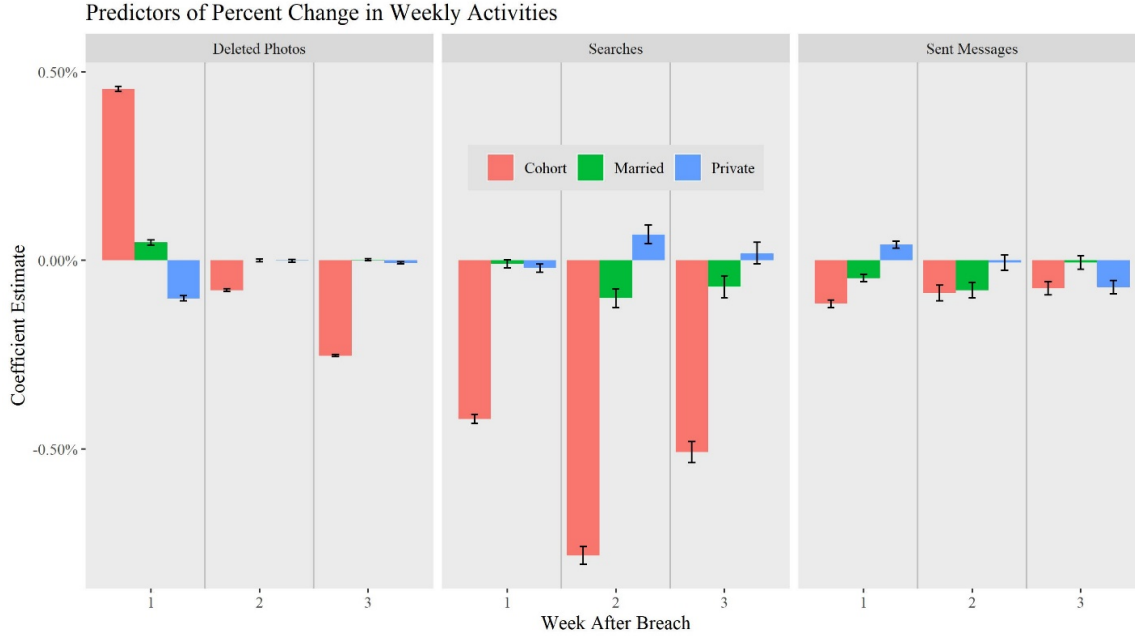


Figure 2.10: Sources of heterogeneity – coefficients and 95% CIs for sources of treatment effect heterogeneity, for each type of activity. Intercept (constant) is added in Figure 9, for ease of interpretation of relative change in activities.

activities. This might be due to “floor effect” – they had more activities to reduce, due to their younger age on the website. For searches and messages, results also suggest that married people reduced more of their messages and searches, compared to single users on the website. In contrast to married users, users who chose to be more private on the website (had no public photo), had smaller decrease (smaller treatment effect) in terms of probability of being active in searches and messages, than those who chose to have public photos on the website.

For deleted photos, married users deleted more photos immediately after the announcement, while private users deleted fewer photos than public users. This might be due to the fact that private users did not understand that being relatively private on a hacked website (by only allowing reciprocal sharing of photos) does not necessarily mean they are protected from other users who might join the website, or from the data breach itself. Being private did not mean they were protected from a data



Figure 2.11: Sources of heterogeneity with main effects – coefficients added to intercept of all DVs (panels) for each week following the data breach.

breach, and in fact, the hackers announced to have hacked the photos as well<sup>15</sup>. This might be an example of a false sense of privacy, where ironically, those who cared most about their privacy and changed their settings to not have public photos available to all, were also the least, and last, to protect their privacy in what can be seen as the only way they could have.

## 2.7 Robustness Checks

In this section, we describe several robustness checks and placebo tests run to assess the “Temporal Causal Inference assumptions” invoked, as well as the validity of the various Causal Inference methods whose results have been presented.

<sup>15</sup>Even though we cannot determine exactly how many photos a user had at the time of breach, prior deletion is a reasonably good indicator of how many photos there are. Therefore, since past behavior ( $X_{it}$ ) was used in the estimation of the treatment effect, we presumably control for the number of photos via proxy information available in the data set.

### 2.7.1 Diff-in-Diff Model

We tested our main results with a “differences in differences” model, as in Janakiraman et al. (2018) and numerous prior studies:

$$y_{it} = \mu_i + \beta_a P_{it} + \beta_{effect} Tr_i \cdot P_{it} + \beta_A A_{it} + \beta_{A2} A_{it}^2 + \epsilon_{it},$$

where  $y_{it}$  is an indicator of whether user  $i$  made any such activity at time  $t$  or not.  $\mu_i$  is the individual fixed effect for user  $i$ .  $P_{it}$  is equal to 1 if the period  $t$  is post-treatment (or lack of treatment, for the control group) for user  $i$ , 0 otherwise.  $Tr_i$  equals 1 if user  $i$  is in the treatment group, 0 otherwise.  $A_{it}$  and  $A_{it}^2$  are the membership age of user  $i$  at time  $t$ , and its square term, to allow for potentially diminishing marginal effect of membership age, as we see in the model-free illustration.  $\epsilon_{it}$  denote error terms, with the usual assumption of being zero-mean Gaussian across users.

Note that adding fixed effects eliminates the need for the treatment indicator  $Tr_i$ , since it is user-specific and therefore correlates perfectly with the individual fixed effects. In order to balance the number of observations for each membership age, the data are restricted to the 8-week window prior to the treatment. Note that this reduction still allows us to gather insights on the full trajectory, since different users are in different membership age categories at the time of the treatment. [All reported results are robust to using the full timeline of all users, which by construction varies between cohorts.]

Since we observe 3 weeks after the treatment (announcement of the data breach) and since we do not want to assume consistent effects for these weeks, we repeat this analysis three times; specifically, for each week after the treatment, we eliminate other weeks’ data. Results of the mean effect on the percent change in number of activities, and of the heterogeneity in users’ response, are reported in Appendix 2.9.2. Results were similar to the results of Temporal Causal Forests. However, there were

some differences between TCF and DID in the magnitude of the changes. We stress that TCF, with its flexibility and nonparametric nature, is more accurate than DID. As mentioned earlier, we verify this claim in a series of simulation studies as will be described in Section 2.7.5, where TCF was found to be more accurate.

## 2.7.2 Generalized Synthetic Control Group

As pointed out by Xu (2017), a major diff-in-diff assumption is that, in the absence of treatment, the mean outcomes of the control group and treatment group would follow “parallel paths”. Though TCF should overcome this by looking for parallel trends across the entire population, and nonparametrically match users in this manner, it is nonetheless important to verify that the main substantive findings are robust. Therefore, we also ran the Generalized Synthetic Control Group (Gsynth) method (Xu, 2017), as a robustness check to Temporal Causal Forests. This method affords several advantages, e.g., it can incorporate multiple time periods following the treatment, allows for heterogeneous effects, and relaxes the assumption of parallel trends between the control group and treatment group by constructing a “synthetic control group” (i.e., a linear combination of trends of multiple users). However, due to computational intensity, all analyses were conducted at the cohort-level, thus allowing for analyses of average effects, and not heterogeneity in effects. We conducted a model selection exercise, choosing among several specifications (adding covariates or not), and among two estimation procedures - either Matrix Completion (MC, as presented in Athey et al. (2021)) or Interactive Fixed Effects (IFE, as presented in Bai (2009)), and present the best-fitting model in terms of goodness of fit in the pre-treatment periods. The model estimated with Matrix Completion is of the form:

$$y_{jt} = \beta_a P_{jt} + \beta_{effect} Tr_j \cdot P_{jt} + \mu_j + \epsilon_{jt},$$



where  $y_{jt}$  is the percentage of active users in cohort  $j$  at time  $t$ . An indicator variable  $P_{jt}$  is equal to 1 if the period  $t$  is post-treatment (or lack of treatment, for the control group) for cohort  $j$ , 0 otherwise. A treatment indicator variable  $Tr_j$  equals 1 if cohort  $j$  is in the treatment group, 0 otherwise.  $\mu_j$  is the fixed effect for cohort  $j$  and  $\epsilon_{it}$  denote error terms, with the usual assumption of being zero-mean Gaussian across cohorts. Results are reported in Section 2.7.3.

### 2.7.3 Bayesian Synthetic Control Method

In addition to Gsynth, we avail of another recently-developed Synthetic Control Method, one from the Bayesian perspective, the “Bayesian Synthetic Control Method (BSCM) (Kim et al., 2020), estimated with a horseshoe prior.

To illustrate the robustness of the results, counterfactuals estimated using both Gsynth and BSCM are presented alongside those of Temporal Causal Forests – removing the average treatment effect on the treated,  $\beta_{effect}$  from the observed behavior of the treated group formed with TCI. Results were found to be robust and comparable to our presented TCF results, for all types of activities, for all weeks after the breach. Since Generalized Synthetic Control Group requires at least 6 weeks prior to the treatment, the results of all three methods include only the cohorts that were at least 6 weeks on the website prior to the treatment, even though in the rest of the paper, when using TCF as the chosen method, we use all estimated cohorts. Interestingly, when looking at Deleted Photos in weeks 2 and 3 after the treatment, the average Control group is slightly increasing the number of deleted photos, relative to what would otherwise have been a downward slope. We do not know of such a reason, but it exemplifies the need to avoid making the parallel trends assumption.

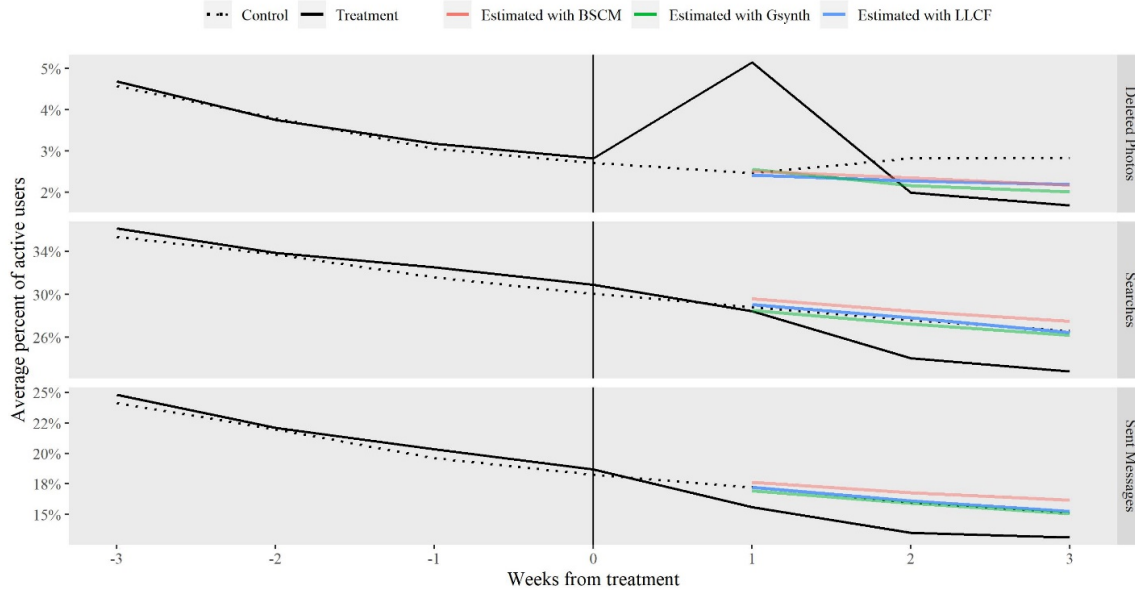


Figure 2.12: Average percent of active users with TCF, BSCM and Gsynth: Control (solid) and Treatment (dashed) group (average across cohorts in each group), three weeks prior and post treatment, alongside the estimated counterfactual number of activities for the control group – estimated either with Bayesian Synthetic Control Method (BSCM, in red) or Generalized Synthetic Congrol Group (Gsynth, in green) or with our proposed chosen method, Local Linear Causal Forests (LLCF, in blue).

#### 2.7.4 Falsification Test: Placebo Test

Possible concerns with Diff-in-Diff methods are explicated by Bertrand et al. (2004). First, as they note, “Obviously, DD estimation also has its limitations. It is appropriate when the interventions are as good as random, conditional on time and group fixed effect.” (Bertrand et al. (2004), pp 250). In the example of an exogenous shock such as the one presented here, the intervention is ‘randomly assigned’ for the simple and tautological reason that all users were affected. However, a possible concern might be that users join the website at different times for reasons that are relevant to the effects of the treatment and are not explained by covariates. In other words, the assignment to cohorts, which is defined as the time of joining the website, might be confounded with the effects of the data breach. Although there is no apparent reason why this might be so, even if there is such confounding effect,

our construction of “Temporal Causal Inference” offers a possible remedy for such a problem: since almost all cohorts are used both as treatment cohorts and as control cohorts for various time periods, whatever possible differences exist will be eliminated or attenuated (at least relative to non-matched reduced-form models). Another assumption that should be tested is the “parallel trend” that underlies the Diff-in-Diff methodology, as explained earlier. It is impossible to test this directly, and therefore we used gsynth and TCF to account for that. In order to further validate this assumption, we ran several Placebo tests, for several weeks prior to the announcement of the data breach. Specifically, we recreated the control and treatment groups via Temporal Causal Inference, with fake “treatments” (i.e., times where no data breach occurred) 3, 4, or 5 weeks prior to the data breach, and reran Temporal Causal Forests with these datasets, on all three types of activities. The analyses showed mainly nonsignificant effects of the treatment, as expected<sup>16</sup>. Due to the nature of these analyses, it allowed us to make further use of the placebo settings in simulation study, as we describe next.

### 2.7.5 Simulation Studies

In order to assure that our chosen method is able to recover the treatment effects even in cases of noise and heterogeneity in treatment, we ran two types of simulation studies, as follow:

#### 2.7.5.1 Synthetic Simulation

The first was purely synthetic, and included data generated using a Diff-in-Diff model, with pre-specified error around the covariates and around the treatment effect. We modeled the log number of activities as dependent variable in all synthetic simula-

---

<sup>16</sup>In “deleted photos”, we saw small, significant, increases in number of activities in weeks 2 and 3 following the “fake treatment”. In these cases, the effect was opposite to that of the data breach. This might indicate that we *underestimate* the effect of the breach.

tions, in order to not have sensitivity to generation of linear probability model with a binary data simulation. We compared both (1) Causal Forests (CF), (2) CF with the Local Linear Correction (Temporal Causal Forests) and (3) the model that generated the data, Diff-in-Diff. Due to the improvement of TCF over CF, we show only TCF in the below plot. Somewhat surprisingly, although DID was the data-generating process, TCF recovers the effects almost as well as DID, in all pre-specified settings (Figure 10, left panel). Moreover, when the individual treatment was correlated with the individual treatment effect, via modeling it to be  $\tau_i \sim N(\mu_i, \sigma_\tau)$ , where  $\mu_i$  is the individual fixed effect, Temporal Causal Forests outperformed DID, in all pre-defined and reasonable variance settings.

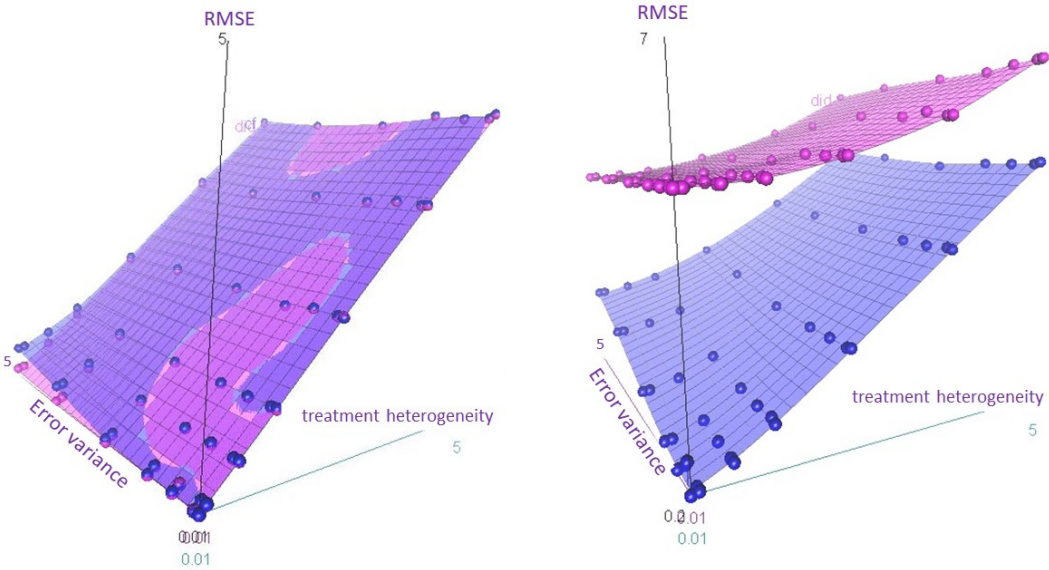


Figure 2.13: Results of simulation studies: RMSE of Diff in Diff (pink) and Temporal Causal Forests (purple) as a function of treatment heterogeneity and error variance. Left panel shows a fit to a simple linear model that generated the data. Right panel shows the fit of a model where the treatment effect is correlated with the individual fixed effects. While in the simple model TCF is able to recover the treatment effect almost as well as DID, in the non-linear model TCF outperforms it (i.e., has lower RMSE) and is able to recover the treatment effect far better for all pre-defined settings.

### 2.7.5.2 Placebo-based (Real Data) Simulation

Due to the nature of the dataset, we wanted to verify that, even with the noise and heterogeneity in activities on the website, we are able to recover a pre-defined treatment effect. In order to do so, we used the placebo setting (pretending the announcement happened before it actually did). In particular, we took placebo of 5 weeks prior to the data breach, then created fake treatment effects of various means and variances. Temporal Causal Inference was able to recover the effects accurately.

Taken together, the robustness checks, placebo analyses, and simulation studies (either fully simulated or placebo-based) suggest that the standard TCI assumptions hold, and that the chosen method – TCF – is able to recover treatment effects both in synthetic data and our real data.

## 2.8 Conclusion and Future Directions

The increase in frequency and severity of data breaches calls for research along several interrelated lines, including prevention, detection, assessment, and *post hoc* remediation. The consequences of data breaches are more than merely financial; they pose individual risk, privacy violations, and loss of trust. In order to help mitigate and assess the consequences, it is critical to understand the range of reactions that data breaches engender. Notification laws were put in place to reduce the risk of financial loss due to such invasions, especially so identity theft. While much research has focused on measuring the effects of exogenous shocks on public companies whose data are widely available, surprisingly little is known regarding individual-level reactions to such breaches, perhaps owing to the need for detailed trajectory data from site users; although such data are rarely made available to researchers, firms track it as a matter of business practice, so could readily avail of causal inference methods in order to assess and mitigate the consequences of a data breach, offer appropriate

compensation, and recover trust.

The construction of *Temporal Causal Inference*, in which the control group was taken to be an older cohort of users, supported the key Causal Inference assumptions, such as un-confoundedness, which is required for measurement. Both average and individual-level effects can be statistically teased out relative to confounds such as typical reduction in number of activities, or differences due to demographic and psychographic traits. Results strongly bear out such differences: married people had more extreme treatment effects than single ones, and private users on the website were less extreme in their changes in activities, than those who were more public on the website.

We must stress that, although the method developed and applied here fully generalizes to assessing other information shocks and data breach incidents, the specific covariate effects almost certainly do not: a great deal depends on the nature of the shock, the individuals compromised by it, and their relationship to the focal firm; for example, marriage is unlikely to be a key demographic implicated in reaction to a data breach in a commercial store setting. However, given that firms typically have a great deal of individual-level information on customer history and demographics, it should be possible for them to paint a rich portrait of the sorts of customers who are *differentially* put off by the breach itself, based on their post-breach usage behavior and prior trends available at large. This is a critical issue in Customer Relationship Management, wherein firms must fashion heterogeneous incentives across the customer base, that is, to offer each customer specific benefits that he or she finds valuable. It may be that, even among customers who react negatively to the breach, some will respond to very different reparative incentives, e.g., some preferring security services (as in the recent Equifax breach), and others financial concessions (as in the data breach to Target; Kude et al. (2017)).

Although we are loath to offer policy implications based on this one study, the

degree of post-breach average activity reduction was surprisingly modest, particularly given the nature of the website and the media storm following the data breach. What appears to be the case is that some users are initially upset and reduced their activity accordingly, but, in relatively short order, we can see a hint of “life returns to normal”. Again, whether this generalizes to other breaches is an empirical question, but it may be that firms should attend to breaches vigorously in the short-term, when users appear to alter their behavior most. That said, policy makers, users of breached websites, and customers of breached stores, would do well to be far more vigilant in the long-term, especially about undertaking actions to both protect their personal data and incent firms they do business with to enact more stringent security measures. Finally, for this one extreme data breach, some users suffered documented psychological distress: self-harm, loss of livelihood, or divorce. While our method can capture a relatively circumscribed range of data breach effects, it is important to underscore that there are far more pernicious consequences. Offering identity protection or best-practices counseling to those at greater risk (e.g., married, sharing more personal information, etc.) may prove effective in reducing such outcomes.

## 2.9 Appendices

### 2.9.1 A1: Heterogeneous Treatment Effects Estimates

#### Heterogeneous Treatment Effects Estimates

	Sent Messages			Searches			Deleted Photos		
	1	2	3	1	2	3	1	2	3
Week	-0.001*** (0.0001)	-0.001*** (0.0001)	-0.001*** (0.0001)	-0.004*** (0.0001)	-0.008*** (0.0001)	-0.005*** (0.0001)	0.005*** (0.0000)	-0.001*** (0.0000)	-0.003*** (0.0000)
Cohort	-0.0005*** (0.0001)	-0.001*** (0.0001)	-0.0001 (0.0001)	-0.0001 (0.0001)	-0.001*** (0.0001)	-0.001*** (0.0001)	0.0005*** (0.0000)	0.0000 (0.0000)	0.00002 (0.0000)
Married	0.0004*** (0.0001)	-0.0001 (0.0001)	-0.001*** (0.0001)	-0.0002*** (0.0001)	0.001*** (0.0001)	0.0002 (0.0001)	-0.001*** (0.0000)	-0.00001 (0.0000)	-0.0001*** (0.0000)
Private	-0.016*** (0.0001)	-0.027*** (0.0001)	-0.022*** (0.0001)	-0.006*** (0.0001)	-0.039*** (0.0001)	-0.037*** (0.0001)	0.028*** (0.0000)	-0.003*** (0.0000)	-0.005*** (0.0000)
Constant	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0000)	(0.0000)	(0.0000)
Observations	49,993	49,993	49,993	49,993	49,993	49,993	49,993	49,993	49,993
R <sup>2</sup>	0.014	0.003	0.003	0.102	0.078	0.025	0.267	0.036	0.415
Adjusted R <sup>2</sup>	0.014	0.002	0.003	0.102	0.078	0.025	0.267	0.036	0.415
Residual Std.Error	0.011	0.023	0.02	0.012	0.027	0.032	0.008	0.004	0.003
F Statistic	233.353***	42.149***	43.784***	1,897.490***	1,402.283***	424.105***	6,080.306***	625.951***	11,806.570***
Note: *p < 0.05; **p < 0.01; ***p < 0.001									

Table 2.2: Heterogeneous Treatment Effects Estimates



### 2.9.2 A2: Differences in Differences Estimates

#### Differences in Differences Estimates

	Differences in Differences								
	Sent Messages			Searches			Deleted Photos		
Week	1	2	3	1	2	3	1	2	3
Treatment Effect	-0.035*** (0.0020)	-0.049*** (0.0020)	-0.046*** (0.0020)	-0.020*** (0.0020)	-0.055*** (0.0020)	-0.059*** (0.0020)	0.019*** (0.0010)	-0.010*** (0.0010)	-0.014*** (0.0010)
After Treatment	0.007*** (0.0010)	0.010*** (0.0010)	0.012*** (0.0010)	0.004*** (0.0010)	0.006*** (0.0010)	0.008*** (0.0010)	0.003*** (0.0004)	0.004*** (0.0004)	0.004*** (0.0005)
Membership Age	-0.042*** (0.0002)	-0.040*** (0.0002)	-0.038*** (0.0002)	-0.041*** (0.0002)	-0.039*** (0.0002)	-0.038*** (0.0002)	-0.012*** (0.0001)	-0.012*** (0.0001)	-0.011*** (0.0001)
Membership Age Square	0.001*** (0.0000)	0.001*** (0.0000)	0.001*** (0.0000)	0.001*** (0.0000)	0.001*** (0.0000)	0.001*** (0.0000)	0.0004*** (0.0000)	0.0004*** (0.0000)	0.0003*** (0.0000)
User Fixed Effects	Yes			Yes			Yes		
Observations	2,770,864			2,770,864			2,770,864		
$R^2$	0.508	0.502	0.497	0.596	0.590	0.586	0.234	0.231	0.230
Adjusted $R^2$	0.438	0.430	0.425	0.539	0.532	0.526	0.124	0.121	0.120
Residual Std.Error (df = 2424502)	0.301	0.303	0.304	0.317	0.319	0.321	0.165	0.164	0.164
Note: *p < 0.05; **p < 0.01; ***p < 0.001									

Table 2.3: Differences in Differences Estimates

## CHAPTER III

# Our Data Driven Future: Promise, Perils, and Prognoses

This essay is a joint work with Fred M. Feinberg. A previous version of this essay has been published in:

**Turjeman, Dana and Fred M. Feinberg (2020) “Our Data-Driven Future: Promise, Perils and Prognoses”. Review of Marketing Research: Continuing to Broaden the Marketing Concept, Vol. 17**

### 3.1 Abstract

Nowadays, most of our activities and personal details are recorded by one entity or another. These data are used for many applications that fundamentally enrich our lives, such as navigation systems, social networks, search engines, and health monitoring. On the darker side of data collection lie usages that can harm us and threaten our sense of privacy. Marketing, as an academic field and corporate practice, has benefited tremendously from this era of data abundance, but has concurrently heightened the risk of associated harms. In this paper, we discuss both the great advantages and potential harms ushered in by this era of data collection, as well as ways to mitigate the harms while maintaining the benefits. Specifically, we propose

and discuss classes of potential solutions: methods for collecting less data overall, transparency of code and models, federated learning, identity management tools, among others. Some of these solutions can be implemented now, others require a longer horizon, but all can begin through the advocacy of Marketing Research. We also discuss possible ways to improve on the benefits of data collection – by developing methods to assist individuals pursue their long-term goals while advocating for privacy in such pursuits.

## 3.2 Introduction

It’s a near-cliché that we live in an age of unprecedented access to data and the decisions they enable. A customer purchasing an appliance, an employer vetting candidates, a suitor seeking a relationship, or a couple picking through restaurants, are each faced with a cornucopia of options, often informed by algorithms attempting to sort through and meaningfully prioritize them. While abundant variety may be overwhelming at times, such algorithmic tools and communication services provide an efficient pathway to what we seek. Compare this with the situation of even a few decades ago, where buying something meant visiting multiple stores, or meeting new people could take weeks of dedicated effort. It’s no wonder society has never looked back, nearly universally embracing the ability to rapidly locate suitable, locally-accessible options. What people often fail to realize is that this array of helpful systems thrives on analyzing not only vast data from an anonymized user pool, but also detailed information on individuals. At the time of writing, worldwide smart-phone users have passed the 2.65B mark <sup>1</sup> (McNair, 2018), roughly half the world’s over-18 population. With a smartphone, tablet, or laptop in hand, humanity’s collective knowledge is just a tap away. Yet this is a two-sided process: each of those taps leaves a detailed, time-granular data trail that collectors can mine to help users

---

<sup>1</sup>[content-na1.emarketer.com/global-digital-users-update-2018](https://content-na1.emarketer.com/global-digital-users-update-2018)

succeed in their goals and plans, but can also store information the users would be reluctant to provide if asked directly (Langheinrich and Schaub, 2018). These data can be used for purposes unaligned with users’ desires, in some cases harming their identity, safety, and sense of privacy. This presents a challenge for policy-makers, firms, and customers, who must balance the undeniable advantages the data-driven era provides with the negative aspects of the its potential downstream uses. The goal of papers in this volume is “Given what we know about Marketing, here’s how to improve some aspect of The World”. Our paper will present something of a cycle, though. We, as marketers, have created and benefited from a Brave New World of data collection. This data-driven era has advanced our society in many domains and aspects of our daily lives. But this abundance of data has revealed a darker side: privacy – the ability to be left alone, unknown, and to have control over one’s own information and reflected identity and behaviors – is vanishing. Data collection on individuals starts literally at birth with government records, and continues as purchases and movements through the world and social spaces accumulate, relentlessly preserved for future access. The tension between the advantages of digital data collection and the liberty to be left unrecorded was noted as early as 50 years ago, in Miller (1969) invocation of the “womb-to-tomb dossier”, wherein any citizen’s life record could be instantly called up, by anyone, for any reason. Stigler (1980) and Posner (1981), writing on the economics of privacy, forecasted the ease and value of acquisition of information (data) and the financial benefits of transferring such data to the highest bidder: benefits that in some cases transfer to customers, through higher efficiency in finding whatever they sought. Yet Miller, Stigler, and Posner could scarcely have envisioned a world where smartphones and embedded devices notate our every move, query, and purchase. We, the Marketers, by aiming to sell smarter, helped forge this double-edged sword: the very algorithms that aid and ease personal decision-making can also channel our data toward less-desired objectives. In

this article, we will present some of the undeniable advances afforded by near-costless, disintermediated data collection, but also its darker shadows: “issues” that arose, as well as potential threats we perhaps haven’t yet been made aware of (but could well have already transpired). We will then review several technical, methodological, and psychological approaches that may assist in maintaining the legitimate gains these data entail, while reducing associated harms. Lastly, we will suggest a stream of thought to encourage the societally and personally positive uses of data. These approaches rely on extensive research in marketing proper, as well as other literature streams.

### **3.3 The Promise and Perils of Data Collection**

The ability to observe people’s behavior across multiple domains in their lives allows us to do more than merely sell them products. It allows social scientists, policy makers, health advocates, and others to learn more about people’s behaviors, even down to their genes, and to enhance their wellbeing – if they so desire. However, this may also lead to unanticipated uses that are not in the best interests of the individuals who simply ‘gave away’ their data. In this section, we will illustrate several of the key benefits of data collection, along with their already-observed hazards, and potential risks that may yet to have surfaced.

#### **3.3.1 Shopping and Search Data**

Marketers have long sought granular data with one overarching goal: to anticipate potential customers’ “needs” and accommodate them with superior products, distribution, and appropriate messages. Before the advent of so-called Personal Computers, this meant collection by large firms in the form of customer surveys, TV ad tracking, and individual store audits, which could be overlaid to achieve a rough match-up between marketing policies and eventual retail performance. All this changed due to the

broad advent of supermarket scanners and panel data: every store purchase made by individual households, along with methods and computational resources to process it (Guadagni and Little, 1983). This early “big data” quite literally revolutionized the practice of marketing, whose effectiveness could be accurately measured in something like real-time. The intervening decades have brought two sequential innovations that have vastly refined the granularity of this measurement: internet shopping and mobile devices. It is possible for marketers nowadays to track, even across devices<sup>2</sup>, every element of the customer experience – ad exposures, searches, page views, clickthroughs, shopping baskets, even brick-and-mortar visits – up through eventual purchase, providing unprecedented refinement of our ability to anticipate and meet customer needs. The spate of purchase-directed search data allows marketers to better target each and every one of us. It may aid us in finding things we didn’t even know we wanted, and to ease the process of locating a good match for our desires. However, this also generated a spiraling number of unsolicited offers to purchase items that we may not want or need (although marketers note an illuminating counterfactual: that, without such targeting, we may receive even more such unsolicited offers, for items that suit us less well). It may also create the feeling of being tracked, leading to negative emotions towards the brand, product, or medium through which the ad was delivered, and may decrease purchase intent (Goldfarb and Tucker, 2011; Kim et al., 2019). Nissenbaum (2009) zeroed in on this phenomenon by coining “Privacy in Context”: when data spill over to other venues, it might violate the sense of privacy. More recently, Kim et al. (2019), in the context of ads on websites, found that customers deem a non-acceptable flow of information as either (1) obtained from another source, or (2) inferred by the website, and not directly provided by the customer. As we will soon discuss, such a non-acceptable flow may involve literal geolocation, and even further through “social space”, in which all our personal interactions through social and commercial media

---

<sup>2</sup>[www.digitalcommerce360.com/2016/11/18/why-retailers-should-track-consumers-across-devices/](http://www.digitalcommerce360.com/2016/11/18/why-retailers-should-track-consumers-across-devices/)

are assembled into a granular private dossier of everything we do, spanning periods of years, and sometimes traded by companies. As detailed in recent investigations<sup>34</sup>, ambiguous privacy policies legitimize tracking, and even trading, all users' viewing activities, demographic data, and even stated political and religious beliefs. This may indeed improve advertising targeting, but might also provide latitude for other purposes unintended – and unknowable – to users.

### 3.3.2 Geolocation Data

The appearance of GPS- and location-based apps have ensured we and our objects are 'never lost'. Their popularity was quick and near-universal: two thirds of US smartphone users use location-based apps, such as Google Maps, Waze, and iMaps, at least once a month (Wurmser, 2018). Hailing apps, such as Uber and Lyft, now commonly used in many countries, benefit from the ability to monitor users' location, matching them with a desired ride almost instantly. Location-based apps allow marketers to micro-target based on one of the most telling factors – exact location – down to the foot. Molitor et al. (2020) found, in a large-scale field experiment, that mobile notifications pushed to shopping mall visitors led to a purchase rate lift of 110% over control group (in the same area, with no such notification). While some argue that such location-based targeting should be seen as a privacy invasion, users-at-large are satisfied when more relevant ads are served up to them based on their interests and opportunity to buy<sup>5</sup>. Despite the vast advantages of the availability and accuracy of location data, they may be misused: publicly available location data from a jogging app, despite being anonymized, was found to reveal US military troops in Syria<sup>6</sup>. Location data posted by users on social media have been used by burglars to target

---

<sup>3</sup>[www.dailymail.co.uk/news/article-6465037/Shocking-extent-big-firms-harvest-data.html](http://www.dailymail.co.uk/news/article-6465037/Shocking-extent-big-firms-harvest-data.html)

<sup>4</sup>[apps.bostonglobe.com/business/graphics/2018/07/foot-traffic/](http://apps.bostonglobe.com/business/graphics/2018/07/foot-traffic/)

<sup>5</sup>[www.aboutads.info/DAA-Zogby-Poll](http://www.aboutads.info/DAA-Zogby-Poll)

<sup>6</sup>[www.washingtonpost.com/world/a-map-showing-the-users-of-fitness-devices-lets-the-world-see-where-us-soldiers-are-and-what-they-are-doing/2018/01/28/86915662-0441-11e8-aa61-f3391373867e\\_story.html](http://www.washingtonpost.com/world/a-map-showing-the-users-of-fitness-devices-lets-the-world-see-where-us-soldiers-are-and-what-they-are-doing/2018/01/28/86915662-0441-11e8-aa61-f3391373867e_story.html)

homes when owners vacate<sup>7</sup>. While most location data are given with at least an implicit consent (anyone can presumably disable location detection on their smartphone<sup>8</sup>), some location data are collected without explicit consent (or are buried in the privacy agreement), and cannot be turned off. For example, telecom companies know your exact location even if your phone is switched off, however old or new it is, as long as it has battery power (and SIM card, if applicable). Recent studies found that half of US cell phone users could be tracked door-to-door, every day, some up to 10 thousand times per day, revealing when they were alone and vulnerable<sup>9,10</sup>. These data are usually handled appropriately, but no law forces companies to not sell them – a \$20B business in 2018 alone – to others, or to use it for any purpose that suits their business model. Such lack of transparency and control over the data that are being collected, and who may access them, lie at the heart of the problem... but also hint at ways to help solve it.

### 3.3.3 Health and Genetic Data

Recent developments in genetic sequencing have made DNA testing widely affordable, with test-kits routinely sold for under \$100. Such affordability allows nearly anyone to locate far-flung relatives, to verify their ethnicity, and to play a role in major health developments. Resulting data have already helped detect genes associated with higher risk of type-2 diabetes (Läll et al., 2017) and Parkinson’s (Nalls et al., 2014), and has the potential to create individualized health treatments and heal society from a malfunctioning gene or genetic disease. It has likewise helped solve dozens of open murder and sexual assault cases around the world, by using non-governmental databases such as GEDMatch (Ram et al., 2018). Despite these

---

<sup>7</sup>[www.digitaltrends.com/social-media/nearly-4-out-of-5-of-burglars-use-social-networks-to-find-empty-homes](http://www.digitaltrends.com/social-media/nearly-4-out-of-5-of-burglars-use-social-networks-to-find-empty-homes)

<sup>8</sup>[www.privacyrights.org/blog/google-tracks-location-data-permission-or-not](http://www.privacyrights.org/blog/google-tracks-location-data-permission-or-not)

<sup>9</sup>[www.wired.com/story/locationsmart-securus-location-data-privacy](http://www.wired.com/story/locationsmart-securus-location-data-privacy)

<sup>10</sup>[www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html](http://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html)



advances, the collection of genetic data has also led to unanticipated consequences. For example, mistaken identifications <sup>11</sup> (also Foeman et al. (2015)), in which ethnicity mixtures are often reported with undue accuracy and confidence, leading many to believe they're not who they are, until a second test proves (sometimes) otherwise. In other cases, people that somehow left their DNA at a crime scene might find themselves in court years later Starr (2016). In the future, insurance companies may use innocuously-collected DNA to identify potential risk of illness, charging a higher premium to that individual, or refusing coverage completely, based solely on a higher predicted risk, going well beyond "pre-existing conditions". A short leap from present technology would enable eugenics without genetic manipulation: match-making services (further discussed below) could allow mate-seekers to prioritize one another based on DNA markers for "desirable" traits, much as they do now on observable ones like proximity or education. Those lacking desired genes might be left un-matched. Due to the nature of genetic data, such outcomes may affect not only people who consent to the usage of their DNA tests, but also to any of their family members.

### 3.3.4 Dating and Relationships

Through all of human history, mate-seeking was mediated by families and local social institutions. The advent of mass transportation and print media widened the circle of potential mates to anyone capable of responding to an ad, but was still a one-off, effortful process. This changed forever in 1995, when the first mainstream online dating site, Match.com, literally opened up the world as potential mates, allowing exchange of information, targeted search, and near-instant communication<sup>12</sup>. In recent years, roughly a fifth of opposite-sex couples and two-thirds of same-sex couples met online, surpassing such mainstays as church and even university as a source of

---

<sup>11</sup>[www.nytimes.com/2018/11/19/magazine/dna-test-black-family.html](http://www.nytimes.com/2018/11/19/magazine/dna-test-black-family.html)

<sup>12</sup><https://www.economist.com/briefing/2018/08/18/how-the-internet-has-changed-dating>

mate-finding (Rosenfeld and Thomas, 2012). This is a boon not only to those seeking mates, but to sociologists and demographers trying to understand this most critical of life decisions (Bruch and Newman, 2018). Yet online courtships have also opened up every phase of potential couples' online interactions to others. Publicly viewable online profiles can be linked forever to the participants, who may not wish co-workers, future spouses, or children to view them, perhaps including exaggerations or outright fabrications. The problem is compounded by the easy exchange of suggestive or even lascivious photos that can, in the wrong hands, go viral. Outside countries with explicit legal protections for LGBT users, dating apps pose serious risks; to illustrate, in 2015, grindr (a dating app geared towards seeking same-sex relationships) started hiding user locations by default in several countries, including Egypt and Russia, where non-heterosexual activity is decriminalized but often persecuted regardless<sup>13</sup>.

### 3.3.5 Social Networks

Social media and networks have changed the ways in which we present ourselves, and in which we define our social connections. Facebook, LinkedIn, Twitter, and Instagram, among others, have burst into and infiltrated our lives in the past decade and a half. These platforms have changed our lives in terms of curation of information and knowledge, creating echo chambers and micro-consensus(es) that enable us to see and hear only those we (and the social network) wish to. Each of these and similar platforms may serve as a different mirror to our personality, but for many, the image reflected can indeed paint an accurate portrait of their beliefs and desires. This also allows us to manage our identity and show the world what we can do, allowing hiring personnel, marketers, and recruiters a global reach and massively improved efficiency. Personal privacy in social networks started off as non-existent by default, and few altered the visibility of their profiles when this functionality was

---

<sup>13</sup><https://edition.cnn.com/2014/12/09/world/africa/egypts-gay-community-living-in-fear>

eventually enabled. Following several incidents and public attention, most social media platforms now offer clearer privacy settings, but it was found that many users never read privacy policies or check their personal privacy settings (Acquisti et al., 2017). In addition, recent revelations suggest that even with stricter privacy settings, third party companies may scrape and collate our data from social networks (Dance et al., 2018). In the past few years, we’ve heard repeatedly about incidents related to socially detrimental uses of such social media data, even when they were not illegally breached. The Cambridge Analytica incident (Confessore, 2018) showed that, even without hacking, social media data could be used in a perfectly legal attempt to sway people’s emotions and voting behavior (Cadwalladr, 2017). This and similar incidents, involving multiple countries and incentives, suggested that targeting certain populations, based on their public (and sometimes) private data, can be used to incite violence and broaden the gap between political factions (e.g., DiResta et al. (2019); Cadwalladr (2017); Caspit (2018)). As we will further discuss in Section 3.4, consistency, transparency, and societal welfare should guide the use and curation of such data.

### 3.3.6 Data Breaches

The above examples – whether they entail perils or positives of data collection – assume that the data are held by someone who acquired them legally, and with some form of user consent (including usage by third-parties, which is legal in most cases). Yet data breach incidents are reported nearly daily, and many – perhaps most – incidents remain undetected or unknown to the public. In 2017 alone, more than 2 billion records have been reported breached (a figure reduced to “only” 1.37 billion records in 2018)<sup>14</sup>. No form of data is immune, and in fact every type of

---

<sup>14</sup>privacyrights.org; The measure here includes data records that were compromised due to security breaches. Possible causes are unintended disclosure, hacking or malware, and physical loss. All compromised records were from businesses, educational institutions, government and military, healthcare providers and nonprofit organizations. In reality, the number should be considerably

data outlined earlier was subject to at least one announcement of a data breach<sup>15</sup>. And this holds aside financial data and identity theft – both much-noted for decades – which are similarly acute. In all such instances of data breaches, companies are encouraged to inform and to help those who were affected, but federal law in the US (and in many other countries) fails to mandate public reporting. Breached data in the wild remain there forever; once out, little can be done to mitigate potential harm. Worse, even when users can take post-breach protective action, most neglect to – the so-called “privacy paradox” (Norberg et al., 2007; Athey et al., 2017) – even when their financial records or others integral parts of their identity might be compromised (Zou and Schaub, 2018). “I have nothing to hide” is a phrase still commonly adhered to; yet even those innocent of wrongdoing can be negatively affected – manipulated, impersonated, or targeted for illegal activity – by exposing their data to shadowy actors. Moreover, even in cases where there is clearly “something to hide”, people fall prey to the privacy paradox and to habituation to instances of data breaches: Turje-man and Feinberg (2021) show that, even in the case of a massive, highly publicized data breach to an extramarital affair-seeking website, some users did not change their patterns of site activity, despite having their full identities breached.

### 3.4 Reducing the Perils of Data Collection

In the previous section we described the tradeoff between the positive and negative aspects of data collection. This tradeoff is here to stay, but a variety of actions can be taken to reduce the risks associated with the compiling of data on individuals and

---

larger; for many of the breaches listed, the number of records is unknown.

<sup>15</sup>Examples: Facebook, October 2018 (social and location data of 90 million users breached); Ashley Madison, July 2015 (full profiles, names, email addresses and more of 37 million affair-seekers breached; data made public in August); Target, December 2013 (credit card and purchase data of 70 million customers); Anthem Data Breach February 2017 (exposed electronic protected health information (ePHI) of nearly 79 million patients); mSpy Data Breach(es) – May 2015, and then another issue in 2018 – both including full records of users who were monitored by the app, including location data, private messages, and more.

their activities. In this section, we discuss methods and tools developed in our field and others, with an emphasis on how actors in and outside Marketing proper can take decisive, leading steps – both short-range and longer-term – to mitigate the perils of a data-driven society, while trying to maintain its primary benefits. Our view is that The Perfect is the enemy of The Good: it is better to start somewhere to reduce the potential for harm than to wait for flawless solutions that may fail to materialize. A graphical overview of our discussion appears in Figure 3.1.

### 3.4.1 Privacy by Default – A Short Term Solution

In the previous section we discussed data breaches, habituation to their announcement, and the privacy paradox. In addition, as shown by Acquisti et al. (2007), few users carefully read lengthy, legalese-laden privacy agreements (a process requiring hundreds of hours annually for anyone so inclined; McDonald and Cranor (2008)), and are thus left unaware of the usage of their data by apps they are using (Almuhimedi et al., 2015), and by other entities who receive or buy their data (Schneider et al., 2017). The recently-enacted General Data Protection Regulation (GDPR) of the European Union introduced “Privacy by Default” – according to which the data of each person are not made available to anyone except that person and the collector, unless the former has actively asked to do so (i.e., via an opt-in basis to share the data publicly or to a third-party)<sup>16</sup>. This mechanism also requires that data will be collected only for the purposes of the disclosed and consented goals; and, once these data are no longer needed, they should be eliminated. In such a scenario, the future will entail less data and a higher cost of acquiring it. Fortunately, no more than a decade ago, much research in Marketing was devoted to solving such “lack of data” problems, due

---

<sup>16</sup>“...by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual’s intervention to an indefinite number of natural persons.” - Article 25(2), EU GDPR, “Data protection by design and by default”

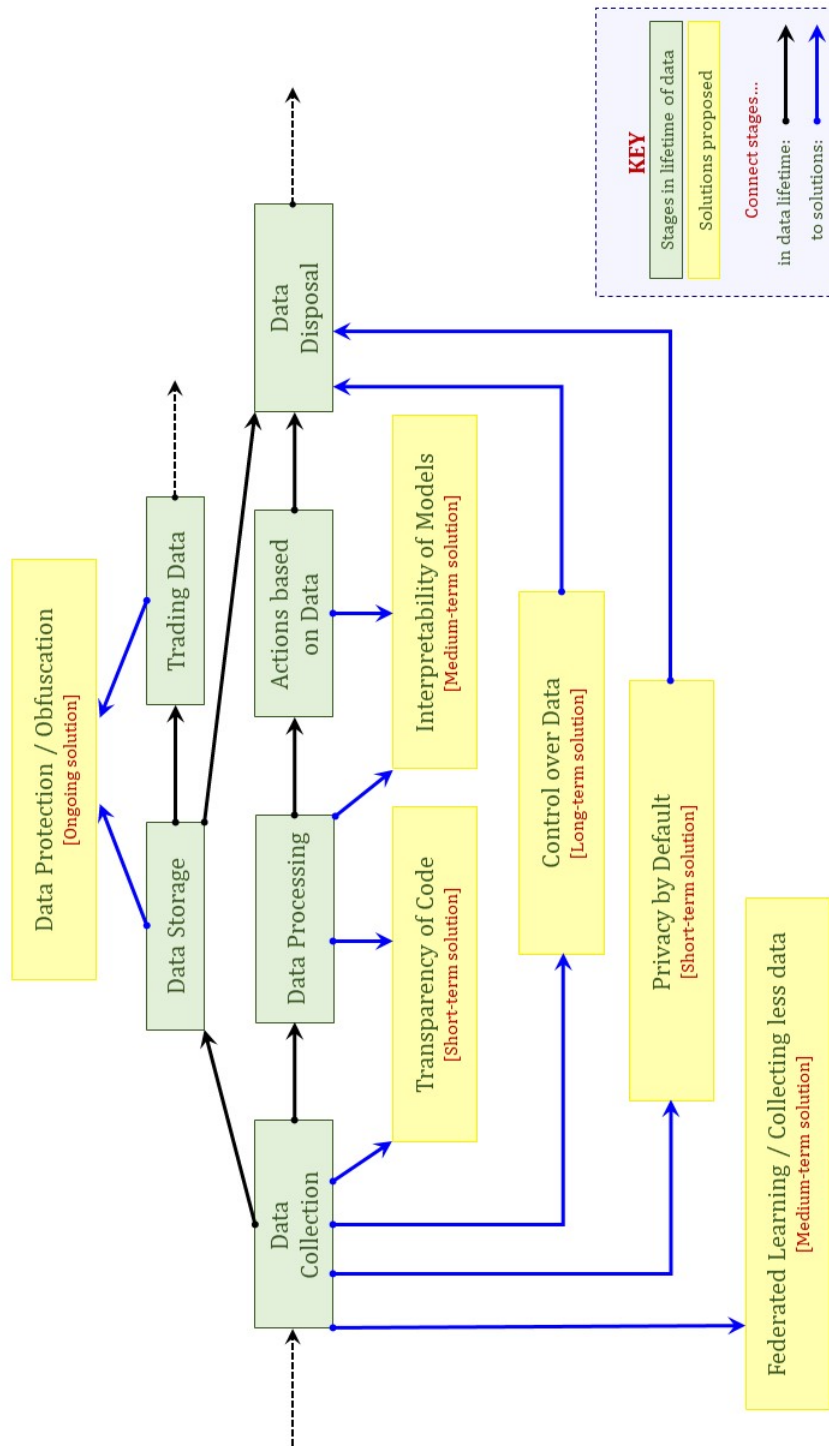


Figure 3.1: Illustration of proposed changes, along with the corresponding stages of data collection.

to historical limitations in acquisition, transmission, storage, and processing. Therefore we, as marketers, are already equipped (and can do yet more) to use the least amount of data to connect customers with their desired products or services. For example, data fusion methods can allow us to collect sensitive data only on subsample of the population, and extrapolate from them to others. Two examples from our own work include Feit et al. (2010), who used a single choice made by car buyers, complemented by prospective customers in a conjoint analysis, to project purchase behavior for cars yet to be manufactured; Turjeman and Tian (2021) are enacting fusion via a variational autoencoder to match an anonymous and sensitive data set to a behavioral, potentially identifiable data set, without endangering anonymity of either side. Schneider et al. (2017) and Schneider et al. (2018) presented Bayesian, model-specific approaches to protect data that are shared with other parties – primarily those that need to be protected due to customer or business privacy. Lee and Anand (2020) propose that data transfer can be limited, and instead models themselves can be transferred, using deep learning techniques. Such methods can preserve desired characteristics of the data without needing the data themselves to be transferred. In computer science, much research addresses data obfuscation, to eliminate the need to save data in raw format (e.g., Bakken et al. (2004); Marino et al. (2019)). In short, foresightful collection and sharing of data can mitigate its potential misuse, legal and otherwise. Advances in marketing modeling can be used to minimize the loss of predictive power that fuels the great advantages of data collection.

### **3.4.2 Code Transparency – A Short Term Solution**

Much of the despair surrounding usage of data concerns the lack of transparency of how they are being used, and by whom. Open-source platforms – software products whose source code is available for all to see, comment on, and (sometimes) obtain permission to change and reuse – may assure customers that there are no misalignments

between the disclosed usage of the data and its actual usage. Obviously, it is easier for a company to protect the usage of its proprietary code when no one sees it, but a sustainable business concerned about customer well-being (and loyalty) can benefit from a publicly-vetted code base. The transparency of code, and usage of open-source platforms, has benefits beyond simply informing users about how their data are being used, but in also improving security of the code itself. Synopsys, a software security and analysis company found (Synoposys, 2015), in an extensive static code analysis, that open-source code had considerably fewer defects (bugs) than commercial software, and was improving on this measure at a comparatively quicker rate. However, as they note, merely using open-source code is not a cure for possible security flaws and bugs if it – and proprietary code that builds upon it – is not reliably maintained (Mansfield-Devine, 2016). In light of the importance of such transparency, several countries, including Canada and the United States, are prioritizing development of governmental code as open-source software (Scott and Rung, 2016; Brison, 2018). Overall, when “done right”, code transparency can improve both the security of our data and our ability to see who is using them and for what purposes.

### **3.4.3 Model Transparency and Interpretation – A Medium-Term Solution**

The advent of deep learning methods, and machine learning in general, has ushered in predictive power and classification accuracy more rapidly than even its early enthusiasts could have anticipated. Relying on such methods, practitioners and researchers alike can enact predictive analyses on vast troves of data, enhance micro-targeting, and improve search results. However, better prediction accuracy usually comes at the cost of reduced transparency of the underlying methodology. With such “black-box” methods, it becomes increasingly difficult to know why a method offers up a specific prediction, which variables were critical to the process, and for what reasons. When



these methods are deployed in innocuous settings, like predicting ad response, they present little risk of harm. But when tasked with predicting likelihoods for illness or recidivism among convicts, baked-in, difficult-to-detect biases might be devastating for individuals. In a landmark case, the state of Wisconsin used a predictive algorithm developed by a private firm – the details of which were a “trade secret” – to help set judicial sentences; challenged because the algorithm was trained on data including race and gender, one defendant sued, escalating the challenge to the Supreme Court, which in declining to hear the case let the use of such algorithms and potentially biased data stand (Simmons, 2017). The potential for serious algorithm-directed error in medicine, insurance, educational opportunity, and legal settings is obvious. More broadly, social scientists need to understand why a prediction was made in order to disentangle the social and psychological bases behind human decision-making. Unveiling why a prediction was made, as opposed to mere predictive accuracy, is a critical area of emphasis. Such an emphasis can help reduce bias, explain why a method predicted what it did, and possibly uncover new social-science phenomena as well as the causal, as opposed to purely predictive, nature of their genesis. As something of an added bonus, knowing which variables are truly useful in prediction (vs. those that are less so) can aid in deciding what sort of data should not be collected in the first place. After all, collecting less data is the best way to protect them from misuse.

#### **3.4.4 Federated Learning – A Medium/Long-Term Solution**

Once data are held by anyone other than the subject, risks of privacy invasion increase. Most advancement in prediction models (e.g., deep learning, as discussed earlier) has required that the data be held by the collector, but this arrangement might not be strictly necessary. With the advancement of mobile devices’ computational power, new machine learning methods can be deployed mainly in the client side (Bonawitz et al., 2017). Google’s AI team is developing such “federated learning”

models, where only updates to the (learning) model are sent, encrypted, to be aggregated with other users' data. Consequently, this maintains and improves the learning of the overall model without ever needing to single out any user's data. Federated learning can vastly reduce amount of data transferred to and stored by firm's servers – most data are held only on the client's side – thereby enhancing data protection. Although such mechanisms are in their infancy, some already argue they might be prone to attacks, if many “bad-actors” coordinate their attempts (Bagdasaryan et al., 2020). Regardless, this approach provides a promising direction in reducing the quantity of data available remotely, and so may increase control over the data shared with service providers while still enjoying the predictive benefits of machine learning and AI techniques more generally.

### **3.4.5 Control Over Data – A Long Term Solution**

As stated earlier, most perils of data collection are due to lack of control over who can access the data, and what are they doing with them. However, controlling the data one gives away, sometimes inadvertently, is no simple matter. Here we propose a somewhat buzzword-laden solution that nevertheless merits our attention: Decentralized Ledger Technology (DLT), such as blockchain or hashgraphs, can be used as identity management systems. DLT represents a technology for sequential, decentralized transfer of information, where every such transfer (or transaction) is recorded and can be both traced and verified. Despite a potentially limitless number of agents in the system, each can create, transfer, and vote for the accuracy and validity of a transaction. Cryptocurrency, such as Bitcoin, is a popular pioneering example of a blockchain implementation, where the commodity (information) that is transferred is a mined “coin”. However, circling back to the underlying mechanism – the mere transfer of information – several recent developments enable identity management using such decentralized approaches, providing a mechanism for controlling who is

accessing which data, when, and why. Such systems have already been proposed to assist in authentication procedures (Lin et al., 2018). Possible extensions may include a “revoke option”, so that if an entity uses the data for a reason not consented to by the user, she will be able to “take the data back”, by changing her public key. While much remains to be done in terms of practical deployment, the general approach can empower and reassure users, and is a particularly worthy topic for academic inquiry.

### **3.4.6 Data Protection – An Ongoing Solution**

The process of protecting data is a never-ending one; every novel protection method is good only until the next clever hacker overcomes it. At that point, further development becomes necessary, requiring frequent security updates. Not collecting the data is the easiest way to protect them, but this comes at a cost of losing the benefits they provide to individuals and society. Some of the methods presented above (data fusion, data obfuscation, federated learning), developed in multiple fields, including Marketing and especially so in Computer Science, can allow us, as a society and profession, to reduce the currently indiscriminate collection and warehousing of data. Further data protection methods, too numerous to list, can and should then be used to protect (or obfuscate) whatever we do deem worthy of collection.

## **3.5 Enhancing The Promise of Data Collection**

Thus far, we’ve discussed the promise and perils of our data-driven future, and laid out several directions to mitigate some of the latter while still benefiting from the former. In this section, we illustrate why we, as marketers, are hopeful about the positive impact that the data-driven era may entail for people’s lives. To a limited extent, this is already happening, in helping us “cut through the clutter” when we have a good idea of what we seek. For example, visitors to Amazon looking for a “food processor” turn up nearly 1000 options; Tripadvisor lets the sushi-craver in Los

Angeles consider 500 establishments; and someone seeking love online in any major city will be confronted with tens of thousands of profiles, sortable by their search preferences. These “proximate goals” – ones easily articulated and whose attainment is verifiable – are ably mediated by recommender systems that help individuals sift the wheat from the chaff. But most true goals in life are evolving and elusive: staying healthy, assembling a rich social sphere, raising successful children, growing wiser. How can information be gathered “safely” to allow individuals to refine and make their way, in the quasi-dark, toward attaining such goals? One possibility lies in data-driven marketing, focused on goals that are hard for us to reach or even elucidate (Dellaert et al., 2018). When guiding a particular individual, algorithms can call on a vast storehouse of information on other individuals: your children aren’t the first whose parents wished the best for them, and no one person trailblazed the search for a life partner. As the information sphere accumulates more and more detail about the choices individuals have already made and paths they have taken, our ability to sort out relatively successful decisions and trajectories from less effective ones not only grows, but allows better tailoring to individuals. Some notable successes have already been realized in individualized medicine. For example, researchers at the Personalized Nutrition Project (Zeevi et al., 2015) found that individual blood sugar response to different foods is highly variable, with the same foods being benign for some and a diabetic danger for others. A key finding was that using both generalized nutritional data and personal microbiome features provides superior glucose response prediction. But to benefit from such findings, individuals would have to allow information about their own biology to be funneled to predictive models created by physicians. One such scenario involves the sort of information systems already commonly used in Health Maintenance Organizations (HMOs) – a patient’s full health records available for analysis – but where users would have control about what additional personal information to share, as well as the power to update, curtail, or delete. While everyone

recognizes the importance of protecting genomic and health information, marketers are often loath to view workaday data on browsing and purchase activity through the same lens. Yet they are enormously hampered by both consumer unwillingness to compromise privacy and the inability to gather data on activities outside their own records and site activity. A proverbial win-win situation could ensue if trusted partners with vetted algorithms could avail of an individual user’s data stream: queries against records that users could control, check for accuracy, limit access to, and willingly enhance to improve their own goal attainment. Such a system could provide nudges (Johnson et al., 2012), alerts, just-in-time suggestions (Nahum-Shani et al., 2017), and – in our view, most importantly – the sort of dynamic “laddering” that allows people to scale long-term peaks step-by-step. For example, someone recently graduating from college may have a vague dream of one day buying a home in a particular city, but no idea what that would entail. Such a goal can become reality using a model based on successful (and less successful) paths taken by others. Then, such data can be used to assist individual users, by providing them information on the types of financial paths that lead to attainment, and the sort of periodic checks, opportunity announcements, and granular feedback that would gently nudge them in the right direction. It is important to realize, however, that data already in the wild are genies permanently released from their bottles. Moving forward, data-driven firms need an incentive to put the power of personal information back in the hands of individual customers, and to assure them that providing voluntary access to data is a net positive for customers and companies alike. Firms, agencies, and governments that can avail of that data without compromising citizen trust are enabled to play a positive role in partnering with individuals to achieve successful, healthy, fulfilled lives. But they can go even further, and do greater good, by considering the goals of society as a whole, and balancing them against immediate individual goals. For example, drivers may all wish to reach a popular venue as quickly as possible, clogging

major roadways with their shared and commonly-pursued goal of minimizing their personal travel time; this data-driven “optimal solution” for each driver may cause all to arrive late, relative to a coordinated one that serves some better than others. How fairness is prioritized over “the greatest happiness for the greatest number” is an open question, and one made more complex when governments (e.g., China’s Social Credit – Liang et al. (2018)), firms, and malicious actors can influence what counts as the common good. Such challenges need to be addressed by sociologists, privacy experts, legislators and, yes, marketers, in the coming decades, as data generation and availability continue their ever-upward spiral.

### 3.6 Conclusion

As we have emphasized throughout, data collection entails both perils and positives. Though we deliberately focused on examples that are relatively easy to place in one of those two categories, other usages of data are considerably more ambiguous. In this article, we detailed the enormous advantages of data collection, as well as its darker shadows. We then presented several solutions that aim to reduce possible perils of data collection, while still maintaining, and even further enhancing, its positives. In summary, marketers and technology experts alike should strive to reduce the amount of identifiable data collected, use them in a transparent way, and protect them, so that harm in misuse and breaches will be, if not eliminated entirely, substantially reduced. In the book “Privacy and Freedom”, public law professor Westin (1968) defined privacy as “*the claim of individuals... to determine for themselves when, how, and to what extent information about them is communicated*”. We must bear in mind that the spate of data we use every day are those of **people**: they are more than consumers, customers, ID numbers, or walking-wallets. These people – we, our families, and our friends – should have control of their “digital identity”.

## CHAPTER IV

# Privacy Preserving Data Fusion

### 4.1 Abstract

Data fusion – the combination of multiple datasets – is a powerful technique to make inferences that are more accurate, generalizable, and useful than those made with any single dataset alone. However, when data fusion involves user-level data, the technique poses a privacy hazard due to the risk of revealing the identities of users. To preserve user anonymity while allowing for a robust and expressive data fusion process, we propose a privacy preserving data fusion (PPDF) methodology based on variational autoencoders (VAE), a nonparametric Bayesian generative modeling framework estimated in adherence to differential privacy (DP) – the state-of-the-art theory for privacy preservation. PPDF does not require the same users will appear in both datasets when making inferences on the joint data, and explicitly accounts for missingness in each dataset by leveraging additional variation in the other to correct for sample selection. Moreover, PPDF is model-agnostic: it allows for inferences to be made on the fused data, without the analyst specifying a model *a priori*. PPDF does so without the original datasets ever coming in contact on a single machine or model. We undertake a simulation to showcase the quality of our proposed methodology, and describe a planned fusion of a large customer dataset from a match making website with a detailed, anonymous survey.

## 4.2 Introduction

Data fusion, or the linkage of multiple data sources, has been applied at leading technology firms, including Facebook (Ryffel et al., 2018), Microsoft (Zheng, 2015) and Google (Papernot, 2019). Data fusion can assist managers to be more informed and more accurately explore customers’ behaviors, preferences and future needs, even when such data are separate. For example, to learn about customer needs through the combination of preference elicitation responses from a survey and eventual purchase data (Feit et al., 2010), or to make more accurate projections of potential market share, through the combination of data on both customers and the general population (McCarthy and Oblander, 2021).

Despite the prevalence and advantages of data fusion, whenever the fused datasets involve any form of customer-level data, the technique poses a privacy hazard of identifying individuals. For example, Sweeney (1997) and Narayanan and Shmatikov (2008) show that a combination of anonymized datasets with other publicly available data can reveal individuals’ sensitive and identifiable information, in a process referred to as “linkage attacks.” The data to be fused, even if they are anonymous or de-identified, might be re-identified with the added data, therefore risking the privacy of the individuals in either of the datasets subject to fusion.

To reduce the risks of identification, while allowing for the advantages of data fusion, we develop a nonparametric Privacy Preserving Data Fusion approach (PPDF). It fuses two or more datasets without the original data sources ever coming in contact on a single machine or within a model, and implements differential privacy, a state-of-the-art framework and methodology to assure privacy preservation. PPDF allows to substantially reduce, or even eliminate, the risk of compromising customers’ privacy and anonymity.

Consider, for example, a company that runs an attitudinal survey on a set of current customers and potential customers, with assurance of anonymity in order to



increase survey response rate and honesty (Bradburn et al., 1979). In addition to the responses from the survey, the company holds other data sources, such as customer relationship management (CRM) or behavioral data. The data from different sources can be fused with the survey data, in order to gather insights on the entire customer population. However, this company might be reluctant to fuse data if it wants to reduce the risk of identifying users as respondents to the survey.

As another illustration, consider two companies who seek to learn from the joint distribution of the combination of their datasets in order to gather insights on market share, complementary purchases, and potential avenues for growth. Each of these companies strives to both protect their intellectual property – the data collected being major part of it – and the privacy of its customer base.

As a final illustrative example, consider a company who wishes to *split* a sensitive dataset into two or more datasets. Such separation of sensitive data can assure that even if a data breach were to occur, the data will be separated and not as harmful as the joint data (the harms of severe data breaches can be dramatically reduced if names, email addresses, and other identifiers would not be stored alongside sensitive choices, attitudinal, and other individual-level data). This company would be able to split the data only if, when an insight on the joint data would be requested, such insight would be possible in a secured manner, through a privacy preserving data fusion. This can become possible using PPDF.

The goal of PPDF is to preserve customer anonymity and intellectual property, while enabling such use-cases and others. Our method is designed to securely combine multiple datasets to gain unified insights, reducing the risk that customers will be uniquely identified in the fused data. We discuss how PPDF methodology may allow companies to collect less data and to store data with fewer identifiers, potentially even splitting datasets to separate locations, and therefore aid in protection against privacy invasions and data breaches.

We exemplify the use of PPDF first by simulation studies, described in Section 4.4. We then illustrate the potential of PPDF by describing a planned fusion of two datasets from an affair-seeking website: detailed and extensive behavioral data (e.g., usage patterns and mate choices – all will be referred to as CRM data) on approximately one million of its customers, and a detailed, anonymous survey, answered by a self-selected sample of approximately 5,500 of the website’s population. The survey includes questions on prior affairs, stated preferences for future affairs, marital status, moral attitudes, vignettes on affairs and other sensitive information. As further described in Section 4.5. The fusion of both datasets may assist us in learning on attitudes and preferences towards affair-seeking, along with eventual choices, but without assurance of privacy, might risk individuals. The proposed methodology will allow us to reduce the risk that individuals’ identities will be compromised in the process. In addition to enhanced privacy, the proposed methodology will allow us to overcome the selection bias inherent in the self-selected sample of survey respondents.

The rest of the paper is organized as follows: PPDF methodology will be detailed in Section 4.3, we’ll detail the privacy enhancement in Subsection 4.3.3.1, and will follow this with a discussion of the types of missing data PPDF can handle, in Subsection 4.3.4. In Section 4.4 we show PPDF’s ability using simulation, and in Section 4.5 we illustrate the data we are planning to fuse. We conclude with a brief summary and a discussion on further directions in Section 4.6.

### **4.3 PPDF Methodology**

Prior work in the domain of data fusion and record linkage ranges from Dunn (1946) – combining population data – to recent advances in combining aggregate and disaggregate data. Record linkage and more complex forms of data fusion have been used in multiple fields, yielding results in economics (Berry et al., 2004), geography (Liu et al., 2020; Dias et al., 2019) and health (Dautov et al., 2019). In the marketing

domain, it has been used to handle missing data in surveys (Bradlow and Zaslavsky, 1999), predict market share (McCarthy and Oblander, 2021), combine choice experiments with CRM data (Feit et al., 2010), enrich parameter estimates and preference predictability (Swait and Andrews, 2003), detect heavy and light users in multiple media platforms (Feit et al., 2013), and to estimate product purchasing and media-watching (Gilula et al., 2006).

Our method, similarly to other data fusion methodologies in marketing, is intended to enhance user- and customer- level data by fusing them with other datasets. However, much prior work in this area have focused on fusing detailed individual (disaggregate) data with aggregate data (Feit et al., 2013; McCarthy and Oblander, 2021). For such aggregate-disaggregate uses, privacy is less of a concern because linkage attacks are not likely, if at all possible, to occur. This is mostly because the aggregate data cannot usually shed light on identities of the people who are in the disaggregate data. PPDF, on the other hand, can fuse data from different sources while protecting individuals' privacy.

PPDF methodology does not require that the same customers appear in both datasets to make inferences on the joint data. Fusion occurs based on the joint distribution of the shared and unique variables, and therefore, under standard assumptions of missingness in the data, to be further described in section 4.3.4, it recovers one dataset's missingness from additional variation made available from the other dataset. Distinguished from prior approaches, the selection bias need not be specified in the model (evidently, there isn't an underlying model specified in the first place). Instead, PPDF recovers the missingness in a nonparametric manner, inspired by advances in Bayesian canonical correlation analysis (Klami et al., 2013; Chandar et al., 2016) and treats each dataset as if it were a random sample from a multivariate random distribution we wish to encode and fuse. Therefore, the marketer or manager who wishes to learn from a survey jointly with CRM data, or from other datasets that

inherently entail sample selection or other missingness, have more opportunity to do so. More importantly, contrary to prior work on data fusion in marketing, PPDF is model-agnostic; that is, it allows for inferences to be made on the fused data, without the analyst needing to specify the model/analysis before fusion.

PPDF methodology also extends the growing stream of privacy preserving methodologies, such as privacy preserving data publication and synthesis (Fung et al., 2010; Takagi et al., 2020; Evans et al., 2020; Ping et al., 2017). Our method builds on differential privacy (DP) (Dwork et al., 2006b), further explained in Subsection 4.3.3.1. DP allows for a pre-specified and context-specific “privacy budget” that can be tuned to the desired risk assessment and tolerated accuracy (or utility) reduction.

Foundational to any data fusion method is the existence of shared common variables. In our illustration dataset, these are common variables such as marital status, gender, and age. Extant data fusion methods usually start with such common variables when matching customers across datasets. However, although matching over shared demographic variables is a natural first step, a well-known result from marketing research is that matching over latent constructs of customer behavior (e.g., preferences, values, and attitudes) elicit improved downstream inferences and predictive accuracy of customers’ needs and desires (see, e.g., Feit et al., 2010). On the other hand, such robust matching on common and latent variables might compromise customers’ identity (within either dataset) and reveal one’s preferences or values along with their identifiable information. This has been illustrated by Sweeney (1997), who relied on demographic data to reveal sensitive health information of public officials in the State of Massachusetts, and by Narayanan and Shmatikov (2008), who relied on inferred preferences when matching de-identified data from Netflix, along with publicly available data from IMDb. The proposed PPDF methodology allows us to learn the joint distribution of both datasets, based on their latent constructs, without compromising anonymity of any user.

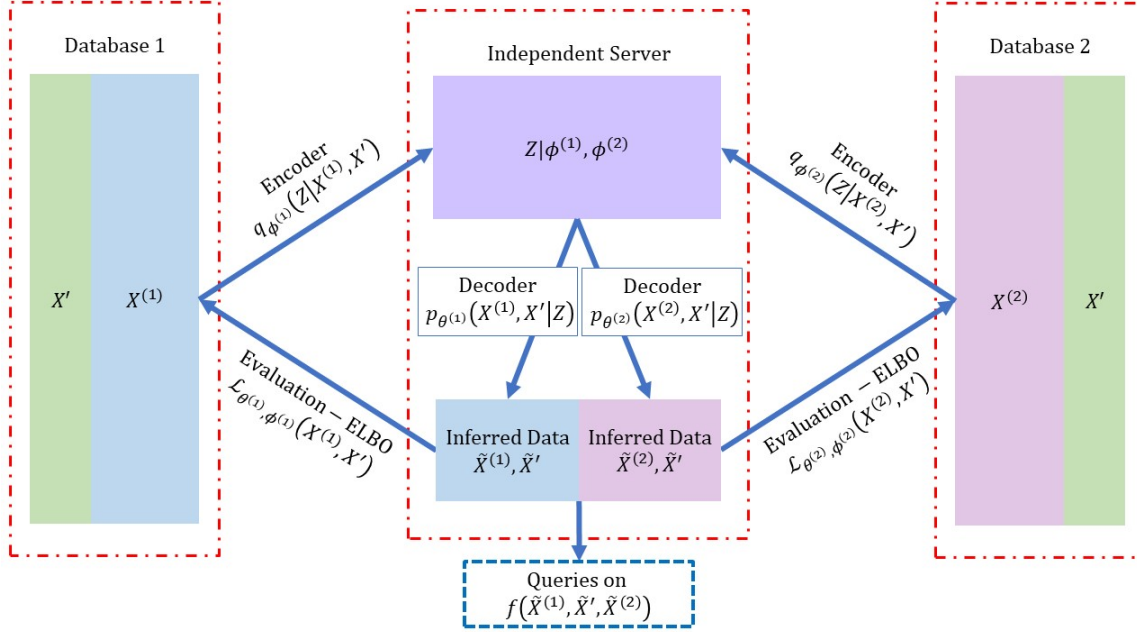


Figure 4.1: Illustration of PPDF of two datasets, each with common variables  $X'$ . Dataset 1 has variables  $X^{(1)}$  and Dataset 2 has variables  $X^{(2)}$ . Encoders  $q_{\phi^{(1)}}$  and  $q_{\phi^{(2)}}$  encode the common and unique variables of datasets 1 and 2, respectively, into  $Z$ . Decoders  $p_{\theta}$  decode the common latent variables into the inferred variables  $\tilde{X}^{(1)}$  and  $\tilde{X}^{(2)}$ , each based on the population of the respective dataset.

Figure 4.1 illustrates the two datasets<sup>1</sup>: Database 1, which is comprised of set of variables  $X^{(1)}$  (e.g., a CRM database including number and type of user engagements, membership duration on the website, contract status, etc.) and common (shared) variables  $X'$  (e.g., age, marital status, ethnicity, gender, etc.); and Database 2, which also includes the common variables  $X'$ , but has  $X^{(2)}$  as its unique variables (e.g., moral attitudes, vignettes on affair-seeking, stated preference for type of affair sought). Importantly, while  $X'$  are common variables in that they have similar structure, they might not be of the same users. In fact, the two instances of those shared variables ( $X'$  in dataset 1 and  $X'$  of dataset 2) might not be drawn from the same distribution – for example, if older people are more likely to respond to a survey – as long as there is sufficient ability to recover the joint distribution. We detail our ability to overcome

<sup>1</sup>In what follows, and for ease of notation, we assume two datasets are to be fused, though this can be generalized into more than two.

selection bias in section 4.3.4.

Given our illustration datasets, or any other data fusion exercise of two (or more) datasets, our goal is to infer their joint distribution (i.e., fused data distribution) while reducing the privacy risks associated with such linkage of datasets. In our exemplary context, we wish to find how the attributes from the CRM database ( $X^{(1)}$ ) *covary* with the response outcomes from the anonymous customer survey ( $X^{(2)}$ ) and explicitly aim to avoid any one-to-one ‘*match*’ between the two datasets. To achieve this, PPDF learns a set of latent representations from the shared and unique variables of both datasets, *encodes* the them, and *transfers* them into the same latent, shared space,  $Z$ . The quality of  $Z$ ’s encoding is evaluated by its ability to *decode* the latent representation  $Z$  back to the original datasets.

Note that the encoding is merely a representation of the joint distribution, does not include the raw data, which might be identifiable, and is differentially private, as will be explained in Subsection 4.3.3.1. Moreover, privacy is preserved since the encoder and decoder, along with the raw datasets, can be on different servers. The encoding and the differential privacy mechanism assure that only differentially private latent representations of the data, and not raw data, are transferred to the common server.

Once the encoder and decoder are optimized (by minimizing the information loss, as will be further described in Subsection 4.3.1), a query based on any subset of variables can be made onto the joint distribution of the remaining variables across both datasets, which is the primary objective of data fusion.

In the following subsections, we will explain the building blocks of PPDF – starting from a single dataset’s encoder and decoder implemented with a variational autoencoder (VAE), improving it through normalizing flow, making it differentially private, and then building the bidirectional transfer learning (BTL) to fuse the datasets.

### 4.3.1 Variational Autoencoders (VAEs)

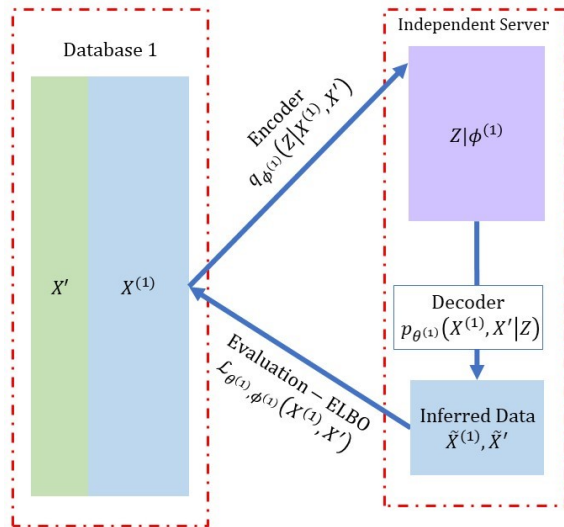


Figure 4.2: Illustration of Variational Autoencoder of a single dataset (without loss of generality, Dataset 1).

Variational autoencoders (VAEs) have been used widely to capture the generative process of images and other data types. In this subsection, we'll describe the variational autoencoders included in PPDF. PPDF comprises of two VAEs, one for dataset 1, the other for dataset 2, that are identical in architecture. Figure 4.2 illustrates a single VAE. For notational simplicity, in this subsection the superscript indicating specific datasets will be suppressed.

A VAE is a self-supervised model that learns the generative model of a given dataset. It comprises two components:

1. An *encoder* (also known as an inference, or recognition, model) that takes the dataset  $\mathbf{x}$  and estimates a set of latent representations  $q_\phi(\mathbf{z}|\mathbf{x})$ , with inference parameters  $\phi$  that capture the data generating process.
2. A *decoder* (also known as amortized inference, or generative model), takes  $\mathbf{z}$  and estimates a model  $p_\theta(\mathbf{x}|\mathbf{z})$  used to reconstruct the original data with set of parameters  $\theta$ , into  $\tilde{\mathbf{x}}$ .

The difference between the original data  $\mathbf{x}$  and the reconstructed data  $\tilde{\mathbf{x}}$  forms the objective we wish to minimize. Through minimizing this difference, the decoder and encoder can *self-supervise* the learning of the dataset’s latent representation  $\mathbf{z}$  and the accuracy of the reconstructed data  $\tilde{\mathbf{x}}$ .

Let  $p_\theta(\mathbf{z}|\mathbf{x})$  be the posterior/decoded latent parameters  $\mathbf{z}$  conditional on data  $\mathbf{x}$ , and let  $p_\theta(\mathbf{x})$  be the marginal likelihood, such that

$$\mathbf{x} \sim p_\theta(\mathbf{x}) \tag{4.1}$$

The marginal distribution, also referred to as the marginal likelihood, is:

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \tag{4.2}$$

Where  $p_\theta(\mathbf{x}, \mathbf{z})$  denotes a deep latent variable model whose prior distributions are flexibly and nonparametrically formed by normalizing flow (more on that in Subsection 4.5). We optimize the variational parameters  $\phi$  such that:

$$q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x}) \tag{4.3}$$

The optimization is done with a loss function, which is derived from the log-likelihood of the data (Kingma and Welling, 2019):

$$\begin{aligned} \log p_\theta(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \left[ \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})} \right] \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \left[ \frac{p_\theta(\mathbf{x}, \mathbf{z}) q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x}) p_\theta(\mathbf{z}|\mathbf{x})} \right] \right] \\ &= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \left[ \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \right]}_{=\mathcal{L}_{\theta, \phi}(\mathbf{x}) \equiv \text{ELBO}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \left[ \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})} \right] \right]}_{=D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}))} \end{aligned} \tag{4.4}$$



We want to maximize the log-likelihood of observing the data. From Equation 4.4, we derived two terms:

1. A latent loss, in the form of Kullback-Leibler (KL) divergence  $D_{KL}$  between the approximate posterior  $q_\phi(\mathbf{z}|\mathbf{x})$  and the actual posterior  $p_\theta(\mathbf{z}|\mathbf{x})$ . The KL Divergence is non-negative,

$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \geq 0, \quad (4.5)$$

and in standard VAEs parameterized to be “close to” the Normal distribution  $N(0, 1)$  in order to keep the divergence suitably small. However, this approximation to  $N(0, 1)$  severely limits the expressiveness of the encoding, and therefore we alleviate this restriction via normalizing flows in Section 4.3.1.2.

2. The *variational lower bound*, or *evidence lower bound* (ELBO)  $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ . Its name derived from the fact that, due to the non-negativity of the KL Divergence, the ELBO acts as a lower bound on the log-likelihood of the data:

$$\begin{aligned} \mathcal{L}_{\theta,\phi}(\mathbf{x}) &= \log p_\theta(\mathbf{x}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \\ &\leq \log p_\theta(\mathbf{x}) \end{aligned} \quad (4.6)$$

#### 4.3.1.1 Optimizing Evidence Lower Bound (ELBO)

Re-organizing equation 4.4 shows that maximizing the ELBO will optimize two measures of interest:

1. Maximize the marginal log-likelihood of  $p_\theta(\mathbf{x})$ ;
2. Minimize the KL Divergence, therefore the encoded approximation  $q_\phi(\mathbf{z}|\mathbf{x})$  becomes closer to the true posterior  $p_\theta(\mathbf{z}|\mathbf{x})$ .

Maximizing the ELBO (or minimizing  $-\mathcal{L}_{\theta,\phi}(\mathbf{x})$ ) will therefore be the objective function with which each of VAEs will be constructed. In practice, this is done by implementing mini-batch stochastic gradient descent (SGD) optimization. The data are split into mini batches of random samples from the original dataset. In each step, the algorithm computes the reconstruction loss on mini-batch  $B = \{x_1, \dots, x_N\}$ , and estimates the gradient  $g_B = \frac{1}{|B|} \sum_{x \in B} \nabla_{\theta,\phi} \mathcal{L}_{\theta,\phi}(\mathbf{x})$ . Then  $\theta$  and  $\phi$  are updated following the gradient direction  $-g_B$ . This will allow the model to get closer to the local minimum of  $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ , thus optimizing the VAE. It is in the SGD that differential privacy will be implemented. However, we first describe the rest of the data fusion process – improvement of VAE using normalizing flows, and the fusion process in the bidirectional transfer-learning phase.

#### 4.3.1.2 Normalizing Flows

One challenge of fitting VAEs is that they are limited in their ability to capture the data generating process. Specifically, VAEs perform encoding using a univariate Normal prior,  $N(0, 1)$ , due to the construction of the loss function (specifically, due to Kullback-Leibler (KL) divergence  $D_{KL}$ ).

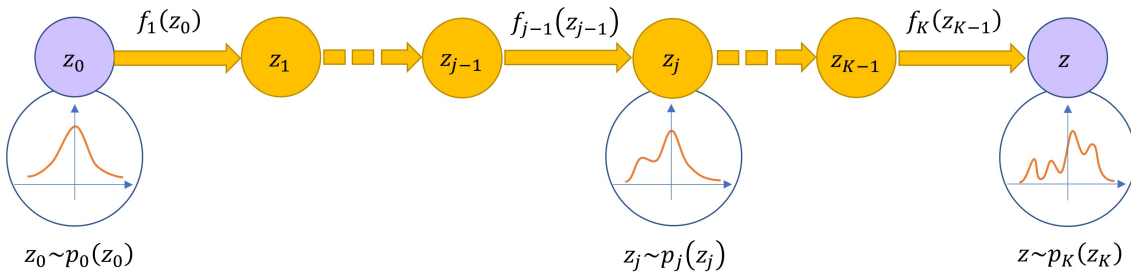


Figure 4.3: Illustration of Normalizing Flow – series of bijective functions  $\mathbf{z} = f_K \circ \dots \circ f_2 \circ f_1(\mathbf{z}_0)$  allows to flexibly represent the data.

To overcome this challenge, we allow the encoder of each dataset to be flexibly formed using a normalizing flow architecture. Normalizing flow (Rezende and Mohamed, 2015; Papamakarios et al., 2019) is a Bayesian deep learning technique that

learns high-dimensional distributions of arbitrary complexity and expressiveness, by undertaking a sequence of non-linear, bijective (volume-preserving) transformations of simpler baseline distributions.

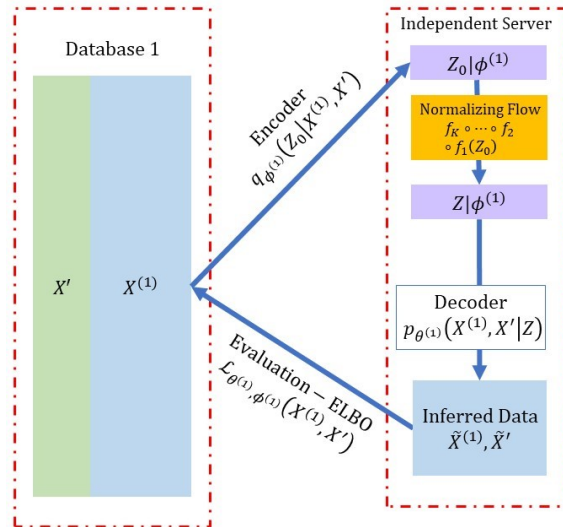


Figure 4.4: Illustration of a single VAE with Normalizing Flow.

Figure 4.3 illustrates a normalizing flow<sup>2</sup>, whereas Figure 4.4 illustrates where normalizing flow will be incorporated: instead of having a simplified latent encoding, distributed  $\mathbf{z} \sim N(0, 1)$ , we add in an intermediate series of bijective functions and a latent parameter  $\mathbf{z}_0 \sim N(0, 1)$ , such that

$$\mathbf{z} = f_K \circ \dots \circ f_2 \circ f_1(\mathbf{z}_0). \quad (4.7)$$

This allows the resultant latent parameters  $\mathbf{z}$  to flexibly capture data relationships of greater complexity. In turn, this enables our encoder and decoder to represent the joint distribution more accurately in the data fusion process.

Figure 4.5 illustrates the improvement of the reconstruction of sample of digits, using normalizing flow. VAEs (Vanilla VAE in this case; Kingma and Welling, 2019) are known to create blurry images (Rezende and Viola, 2018), due to the limitation

<sup>2</sup>Illustration in Figure 4.3 is inspired by Lilian Weng: <https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html>.
















Original Image	Vanilla VAE	NF + VAE
		
		
		
		
		

Figure 4.5: Results of VAE with and without Normalizing Flow: results of VAE (middle column) and VAE with Normalizing Flow (VAE + NF, right column). VAE creates a blurry image, whereas VAE with NF is much clearer.

described in Subsection 4.3.1. The use of a normalizing flow (in this case  $K = 7$  functions) allows a richer underlying regularization distribution, and results in much clearer images, that capture the original digits well.

### 4.3.2 Bidirectional Transfer Learning

In the previous subsections, we discussed the creation of VAEs. In PPDF, we create a single VAE for each dataset we fuse. This explanation omitted an important part of the process: the data fusion itself. If each dataset is encoded, where does “the magic of data fusion” occur? In this subsection, we explain the process of bidirectional transfer learning (BTL).

Based on the conceptual framework of Bayesian canonical correlation analysis (BCCA), we’re treating each dataset as a multivariate random variable with unknown parameters. The encoders – with the flexibility of normalizing flow – are encoding

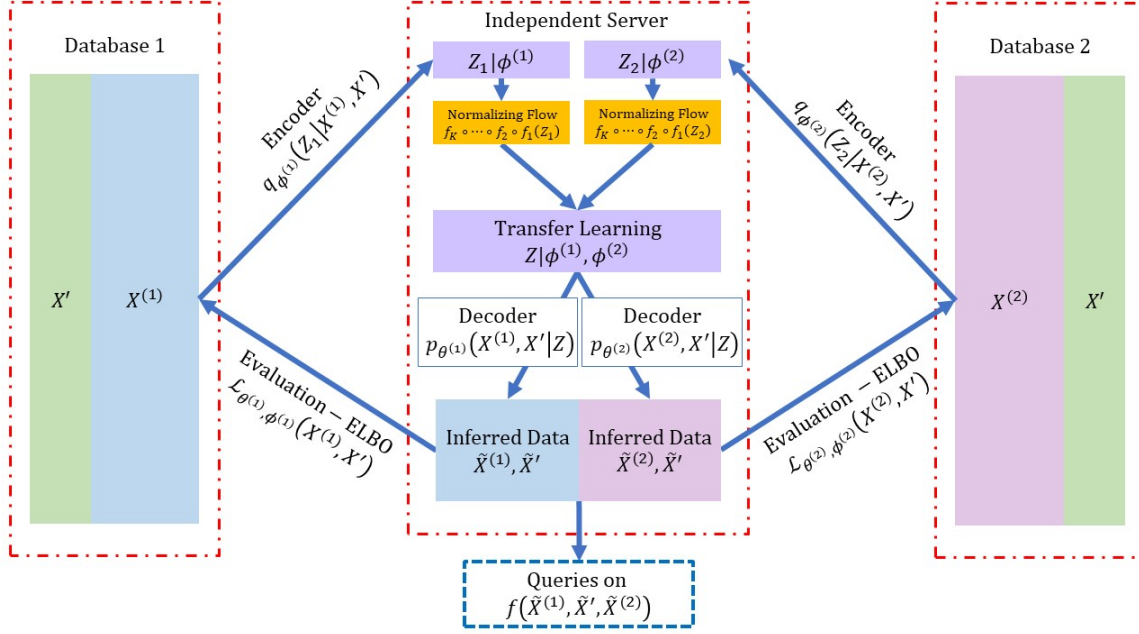


Figure 4.6: Detailed architecture of PPDF – two VAEs and each with its own normalizing flow.

each dataset into latent representations. This representation is of both the common and unique variables of each dataset. The common variables will allow us to construct the mapping between the two datasets in a formation of a joint latent representation. Figure 4.6 illustrates the full PPDF architecture.

Consider  $Z_1$  and  $Z_2$ , the latent representation of datasets 1 and 2 respectively, encoded using parameters  $\theta^{(1)}$  and  $\theta^{(2)}$ . They iteratively go through a bidirectional transfer learning, such that  $Z$  is constructed using both datasets. The training is optimized based on three losses:

1. Minimize the self reconstruction error (through maximizing the ELBO, see Subsection 4.3.1.1). This will minimize the difference between the inferred (decoded) data and the original data.
2. Minimize the cross-reconstruction error. I.e., minimize the error of reconstructing dataset 1 from dataset 2, and vice versa, through the usage of the common variables.

3. Maximize the correlation between the latent representations  $Z_1$  and  $Z_2$ .

Using the augmented variational parameters, data fusion (i.e., cross-reconstruction) can then occur as probabilistic imputations from a joint posterior predictive distribution. In other words, by garnering the posteriors  $p_{\theta^{(1)}}(X^{(1)}, X'|Z)$  and  $p_{\theta^{(2)}}(X^{(2)}, X'|Z)$  - where  $Z$  are the common latent representations,  $X^{(m)}$  are the dataset-specific variables, and  $X'$  are the shared variables - we can obtain the posterior predictive joint distribution,  $f(X^{(1)}, X^{(2)}, X')$ .

Equipped with  $f(X^{(1)}, X^{(2)}, Z)$ , the end result of this data fusion is that, for every entry in either dataset, we have a probabilistic reconstruction of the matching entry from the other, such that with  $p(x_i^{(1)}|x_i^{(2)})$  we can construct  $[x'_i; x_i^{(1)}; x_i^{(2)}]$ .

As another conceptual illustration, our combination of VAE and BTL can be thought of as a ‘game of charades’ between two agents (i.e., a federation of models) who must recreate the information held by the other through iteratively providing one another with a set of ‘clues’ from the original raw data. As a rough illustration of a single pass of the VAE for our context, first, the *encoder* agent encrypts the set of unshared, or unique, attributes from one dataset into a set of latent variables. These are then passed to the decoder model, alongside the shared common attributes (e.g., demographics, usage metrics) that the unshared attributes were originally indexed to, which serve as the ‘clues’. The *decoder*, with learning reinforcement from the encoder, then engages in two objectives: to decipher and reconstruct the encoder’s abstract latent variables into their original format, and to learn the relationship between the reconstructed unshared attributes and their associated shared common attributes. For the first task, the decoder iteratively provides reconstruction guesses, as the encoder’s role is to confirm their degrees of accuracy. If the guesses enter into a tolerable range of accuracy, the decoder imputes the joint distribution of the shared common attributes to: (1) reconstructed attributes from the encoder’s dataset, e.g., the anonymous survey, and (2) unshared attributes of its “own dataset”, e.g., the

customer database. What this begets then is that when given a set of shared attributes – such as a set of customer segmentation variables-of-interest – the decoder now possesses the ability to generate the associated values from both the survey and database, and thus, complete the fusion.

Using the combination of these approaches, our framework is capable of fusing two datasets into a single joint outcome. Importantly, it does so without the original datasets ever coming into contact on a single machine or within a model, thereby reducing the risk of compromising customers’ privacy and anonymity. In the next subsection, we’ll plot differential privacy ability within the process.

### **4.3.3 Privacy Preservation Measures and Controls**

One of the essential parts of the proposed methodology is the ability to preserve privacy. There is an inherent tension between privacy and accuracy, when it comes to fusing datasets. The best data fusion will match each user’s variables in a dataset to the same user’s variables in the other dataset. However, such unique identification might reveal the user’s full set of attributes, and in some cases will allow the researcher to uniquely identify them along with traits they did not choose to disclose, or that can potentially hurt them. This risk is known as a “Linkage attack,” and has been demonstrated in Sweeney (1997) and in Narayanan and Shmatikov (2008).

On the other side of the privacy-accuracy trade-off, a completely private data fusion might take merely the summary statistics of the variables of the entire population in one dataset, and correlate those with that of the other dataset. This will allow learning the joint distribution on the entire population, but will not allow for heterogeneity and covariance across datasets.

Consider the behavioral CRM dataset from an affair-seeking website, along with detailed, anonymous survey responses on attitudes, past behaviors and demographics. Any identification that will result from the data fusion might harm individuals in the

datasets, who may have wished to stay anonymous and not reveal their attitudes or behaviors. Therefore, in such cases, the data holders may choose to prefer privacy over accuracy. As another illustrative example, less privacy might be deemed necessary when handling datasets from public sources, since it is reasonable to assume that individuals in such datasets are not expecting privacy guarantees, by the mere presence of their data in a public dataset<sup>3</sup>.

Therefore, it is up to the data holders to assure they're in line with customers' expectations, regulations, privacy policies and known risks, when using the proposed methodology, or indeed any data fusion method. The sensitivity of the data, along with the sensitivity of the usage of the fused data, should guide in deciding on the level of privacy guarantees.

As part of the proposed privacy preserving methodology, we offer tuning mechanism that will enable the data holder(s) a higher sense of control over the level of privacy vs. accuracy. This tuning mechanism is achieved using differential privacy.

#### 4.3.3.1 Differential Privacy

Differential Privacy, first introduced by Dwork et al. (2006b), is a mechanism used to formalize the trade-off between privacy and accuracy through the introduction of added noise. It allows the researcher to tune the risk associated with identifying a person from a database, and explicitly set a “privacy budget.” Differential privacy is considered state-of-the-art among current privacy preserving methodologies, and has been used for data publication or data release (Takagi et al., 2020; Fung et al., 2010), including the release of data from the 2020 Census<sup>4</sup>.

Other privacy preserving methodologies, such as K-Anonymity (Sweeney, 2002) (obscuring the data such that every person cannot be distinguished from other  $K - 1$

---

<sup>3</sup>While it is reasonable to assume that privacy expectations are low, the researcher might still want to err on the side of caution, and choose to de-identify individuals.

<sup>4</sup><https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html>



people in the dataset) and L-diversity (assuring that each variable has at least  $\mathcal{L}$ -well-represented values) have been proposed to enable data publication and data synthesis. While such methods may assist with relatively small number of attributes, they fail to scale to large datasets, and might still suffer from various privacy attacks that will reveal identities of the people represented in them (Li et al., 2007; Domingo-Ferrer and Torra, 2008). Nevertheless, they have been found suitable for multiple uses, most notably password checkup tools such as “Have I been Pwned” And Google’s security checkup (Li et al., 2019). We rely on differential privacy due to it’s ability to withhold to richer datasets and due to the mathematical guarantees and clear tuning parameters it enables.

Differential privacy relies on the assumption that if it is impossible to ascertain that any particular user’s data were used in an analysis, their privacy is preserved. From another angle, with a differentially private algorithm, each individual in any given dataset has a bounded probability to be revealed as included in the dataset, relative to another dataset that only differs in the removal of their data. Differential privacy therefore relates to the to assurance (up to a bounded probability), that for the inclusion of an individual in a dataset would not change the outcomes relative to a dataset that does not include their data.

We first begin by defining  $\epsilon$ -differential privacy. Consider two adjacent datasets,  $D$  and  $D'$ , that are the same except that dataset  $D'$  has one more observation, i.e.,  $D' = D \cup x_i$  where  $x_i$  are the data of individual  $i$ .

An algorithm  $M$  is considered  $\epsilon$ -differentially private ( $\epsilon \in \mathbb{R}$  and small as desired), if for every output  $S$ , we receive the same output  $S$  with the other dataset  $D'$  at a probability that is at most  $e^\epsilon$  that of dataset  $D$ . A low  $\epsilon$  means that for the two datasets that differ only in the existence of  $x_i$ ’s data, we have very low probability of distinguishing any given output. This makes inclusion of  $x_i$  in the data to be very

hard to detect:

$$\Pr(M(D) \in S) \leq e^\epsilon \cdot \Pr(M(D') \in S) \quad (4.8)$$

This can also be seen as: person  $i$  cannot be revealed as a respondent to a survey, if they haven't responded to it. The probability of being identified as a respondent, through a variation in the outputs, would be very low in such case. A differentially private dataset would allow us to state that even if  $i$  is a respondent to the survey, the probability of them being identified as a respondent is very low as well; it is at most  $e^\epsilon$  more likely.  $\epsilon$  is therefore a measure of the "Privacy Loss", and, by construction, smaller values of  $\epsilon$  would lead to lower privacy loss.

As another variation of differential privacy, Dwork et al. (2006a) added an upper bound of the individual risk  $\delta$ , such that:

$$\Pr(M(D) \in S) \leq e^\epsilon \cdot \Pr(M(D') \in S) + \delta \quad (4.9)$$

The addition of  $\delta$  acts as a "failure probability", and should preferably be set such that  $\delta < \frac{1}{|D|}$ . This failure probability acts as a tolerance to the risk associated with identification: allowing for the possibility that  $\epsilon$ -differential privacy is broken with probability  $\delta$ .

We implement  $(\epsilon, \delta)$ -differential privacy, in the variational encoder of PPDF. In particular, we follow Abadi et al. (2016) and implement a differentially private stochastic gradient descent (DP-SGD) estimation of the VAE.

At each step of the training of the VAE, we compute the gradient of the Loss function  $g(\mathbf{x}) = \nabla_{\theta, \phi} \mathcal{L}_{\theta, \phi}(\mathbf{x})$ , or, for a random subset of samples (mini-batch)  $B = \{x_1, \dots, x_N\}$ , compute the gradient of the mini-batch:  $g_t(x_i) = \frac{1}{|B|} \sum_{i=1}^N \nabla_{\theta, \phi} \mathcal{L}_{\theta, \phi}(x_i)$ . The parameters are then updated following the gradient with learning rate  $\eta_t$ , such that the updating of parameters  $\theta, \phi$  is  $\{\theta, \phi\}_{t+1} = \{\theta, \phi\}_t - \eta_t \cdot g_t$ . This is a procedure common for every VAE, but DP-SGD adds two more steps in the computation of the

gradient, to assure privacy.

1. Clipping the norm of each gradient  $g_t$ , in order to assure that the information of each individual in a mini-batch is limited:

$$\bar{g}_t(x_i) = \frac{g_t(x_i)}{\max\left(1, \frac{\|g_t(x_i)\|_2}{C}\right)} \quad (4.10)$$

2. Adding noise from a Normal distribution such that

$$\tilde{g} = \frac{1}{|B|} \left( \sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbb{I}) \right) \quad (4.11)$$

The parameters for the clipping of the norm and for the added noise are computed based on the desired  $\epsilon$  and  $\delta$ , in a process referred to as “privacy accounting”, detailed in Abadi et al. (2016).

#### 4.3.4 Handling Missing Data and Selection Bias

Inherently, every data fusion task is intended to impute missing values: the researcher is imputing the unique variables from one dataset into the other, relying on the common variables in both.

In addition to such missing variables, in almost all realistic datasets, there is a need to overcome the problem of missing values within variables. Missing values in data is a common problem in social science, and in particular, in marketing research. Collecting data from human subjects is highly likely to result in missing information. This occurs for a variety of reasons: unwillingness of users to respond to some questions (Bradburn et al., 1979); changes in experimental design over time, which might result in missing observations for whole variables (Graham, 2009), and flaws in data collection carried out in field settings.

Past techniques for handling missing data involved either removing observations

if they had even one missing variable, or completing the missing data with the observed sample mean (Graham, 2009). These techniques have been shown to be both inefficient and inaccurate. One drawback is that they reduced the sample size, thus possibly creating a non-random sample. They could also lead to inaccurate inferences due to the fact that using the mean of an explanatory variable can change the impact on the explained outcome.

Two common methods are used to handle missing data in the social sciences in general, and in marketing research in particular. The first group is Multiple Imputation methods; key examples appear in Little and Rubin (1989) and Kamakura and Wedel (1997). The second group of methods for handling missing data are Maximum Likelihood methods. Such methods can be based on classical maximization of a model likelihood, such as in Kamakura and Wedel (2000), or on stochastic simulation, such as a Bayesian estimation used in Feit et al. (2010).

Qian and Xie (2014) proposed a Bayesian approach for completing missing data in regression covariates. A major contribution of their work is the ability to derive the missing values and regress over all data, without the need to specify the exact distribution for each covariate, and the relations among covariates. This technique can handle high-dimensional missing covariate problems. However, when there is insufficient information to recover the underlying model, and when there is insufficient data or a too complex problem to handle, this method may not be suitable.

There can be several types of missingness that should be acknowledged and properly handled. Specifically, Rubin (1976) classifies three mechanisms of missing data: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR, also known as non-ignorable).

In the case of MAR, while some of the data are missing, the missingness can be overcome by other observed variables. This means that the cause of missingness may depend on other covariates that are in the data, but not on any unobserved data. A

simple example can be responses for an income question on a survey. People might be reluctant to fill in their income, if they feel that they have too high or too low of an income. However, missing answers can be imputed using a combination of other variables such as level of education, living area, age, etc. Though imputation will probably not retrieve the true individual responses, inferences drawn on large data samples will not be affected by the missingness.

MCAR can be considered a special case of MAR. As implied by the name, the missingness is completely random, and does not depend on either observed or unobserved variables. One possible example is when some of the data are corrupted due to technical error. Another example is when some respondents simply forget to answer certain questions in a randomly-ordered questionnaire. As a more poetic example, consider a dataset on a piece of paper that was under surprising rain. Some observations might have been deleted in random. No worries! Since they were randomly missing, any inferences regarding the data as a whole will be correct<sup>5</sup>, thus representing the true underlying data generating processes, with or without completing the data.

The MNAR (also known as NR, nonrandom, or non-ignorable) missingness mechanism occurs when some of the values are missing and their missingness depends on unobserved data. Therefore, some information in the missing data depends on the missing values themselves, and cannot be fully rectified based on available information; as such, inferences made based on such data might be biased by the missingness. Consider our example from above regarding missing income: if people are reluctant to respond to an income question because they are concerned about scammers, but such concern cannot be explained by any data available to us, then the missingness is non-random, but would revert to MAR if we could somehow account for this missing piece of information, perhaps through a survey question about this particular concern.

---

<sup>5</sup>With a caveat: the standard errors are likely to become larger, since we have fewer observations

In our method, we will allow the data to have missingness of types MAR or MCAR. We will nonparametrically complete (augment) a latent representation of both within-variable missingness and the obvious whole variable missingness, with observed variables from either dataset. We define common variables as those that appear in both datasets, and are measured on the same scale. By semi-common variables, we refer to variables that relate to the same underlying information, but are measured differently. For example, date of birth (DOB) vs. age group. While DOB is a point measure, precise, and can potentially ease the process of identifying a person, age group can be in categories of (say) five years, which makes it harder to identify any one person, even in a relatively small samples.

Given that missingness can occur across multiple data variables, a key limitation to conventional data imputation methods is that model complexity scales with the number of missing values. VAEs overcome this limitation by treating missingness as arising from a single generative model, and instead seek to encode the joint data generating process as a nonparametric random function that may in turn then be used to decode missing values where missingness may arise. The method allows for data – either individual values or entire covariates – to be missing at random (MAR), and for truncation into categories if data are semi-common.

Moreover, if data are MNAR, but the missingness can be accounted for (becoming MAR) using the other dataset, PPDF will be able to account for it as well. Consider a self-selected survey as our illustrated dataset. If older users were more likely to respond to the survey for some reason, such selection might bias our imputation of the age variable and variables that correlate with it, such as times married or number of affairs the respondent had. With BTL, PPDF will be able to overcome such missingness, as long as the common variables bridge the missingness. So, if one dataset had missingness not at random, using the common variables (e.g., age) and the other (sufficiently informative) dataset, we can potentially overcome this

missingness, essentially making MNAR on a single dataset become MAR in the joint dataset.

## 4.4 Simulation Studies

After going through the architecture of PPDF, we will now showcase its ability to fuse datasets for which we know their underlying joint distribution. The first simulation is based on MNIST data. Described in Deng (2012), the MNIST dataset is frequently used to assess classification methods. It includes 60K black and white images of numeric digits, each with  $28 \times 28 = 784$  pixel. In Figure 4.5, we exemplified the improvement of VAE+NF over VAE in the reconstruction of the digits.

To illustrate PPDF, in the next simulation, we will split each image into two – allocating a portion of middle pixels as if they were common variables – and will then fuse them back.

Figure 4.7 shows the reconstruction loss of the fusion of MNIST digits, with varying levels of added noise  $\epsilon \in \{0.5, 2, 8\}$ , where smaller levels of epsilon means larger noise added to the DP mechanism. It is clear that adding more noise dramatically increases reconstruction loss in recovering the MNIST digits. It is also apparent that in this simulation, the added noise also makes the training go in the wrong direction (for example, the increase in reconstruction loss of both the training and test data, when large noise is added, at epochs 13-15). This is due to the added noise that might make the gradient not be sufficiently informative to approach the local minimum.

In order to highlight the role of  $\delta$  and  $\epsilon$  in varying the noise levels of the resultant fusion, Figure 4.8 shows the reconstruction loss of the last epoch (in this experiment, 10 epochs), for varying levels of noise. In this simulation, we used MNIST data, where each digits is comprised of  $28 \times 28 = 784$  pixels. We left 300 pixels from the center of each digit to be the common variables for each observation, and  $\frac{784-300}{2} = 242$  pixels were considered unique variables for each dataset. The smallest added noise was no

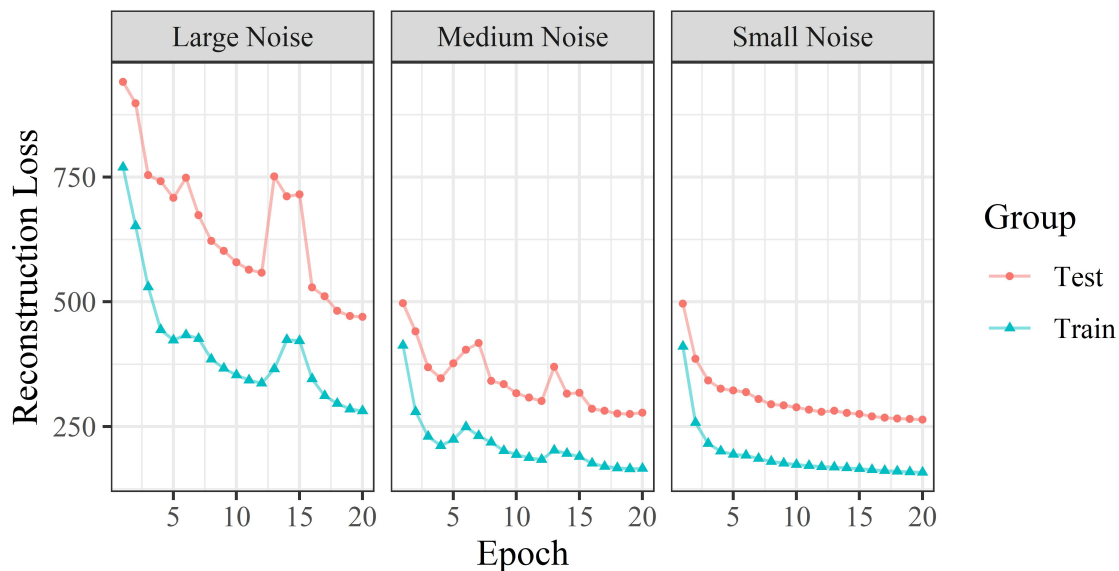


Figure 4.7: Simulation results on MNIST dataset – reconstruction loss (the loss relative to the original images) as a function of epoch – the learning iteration, pre-set to 20 iterations. Each panel corresponds to a different level of added noise – larger noise means smaller level of  $\epsilon$ . Large noise in this simulation corresponds to  $\epsilon = 0.5$ , Medium:  $\epsilon = 2$ , Small:  $\epsilon = 8$ . In this simulation,  $\delta = 10^{-5}$  and is constant. Learning rate is  $\eta = 10^{-4}$  and common variables are the middle 150 pixels of the 784 pixels in each digit.

noise at all (in Blue line), and acts as a reference. As we vary  $\epsilon$  from 100 to 10 and to 1, we are limiting the privacy budget – increasing the privacy, and adding more noise. This results in higher levels of reconstruction loss. Similarly, as we vary  $\delta$ , we add more noise, which results in higher loss.

Following the fusion, we get the reconstructed images presented in Table 4.1: upper row corresponds to the basic data fusion model with no added noise, and the rest of the columns show the resultant images with the varying levels of  $\epsilon$  and  $\delta$ .

As in any data fusion, the ability to reconstruct depends on the number of common variables. This is usually a given, but if company wishes to split a dataset in order to protect its customers, it can potentially control the number of common variables before splitting. Figure 4.9 show the ability to reconstruct images with varying number of common variables (pixels in MNIST images). The more commonality there is



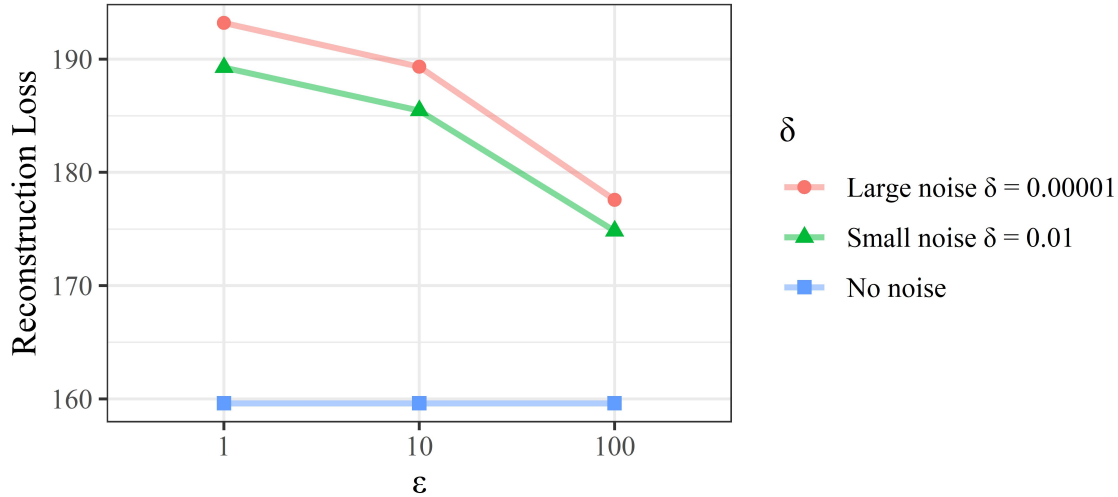


Figure 4.8: Simulation results - varying  $\delta$ : reconstruction loss (the loss relative to the original images) as a function of  $\delta$  and  $\epsilon$ . Each line corresponds to a different level of  $\delta$ , where larger noise (smaller  $\delta$ ) means lower tolerance to re-identification. Large noise in this plot corresponds  $\delta = 10^{-5}$ . Small noise:  $\delta = 10^{-2}$ . No noise means no privacy guarantees at all, and acts as a reference. Learning rate is  $\eta = 2 \cdot 10^{-4}$  and common variables are 300 pixels of the 784 pixels in each digit.

between datasets, the better the reconstruction, and the faster the model runs. Data holders who wish to split datasets may try their split (vary not only the number of common variables, but also which variables remain common) and test the reconstruction loss with varying  $\epsilon$  and  $\delta$  as well, to test the privacy measures for the specific context.

Beyond the tuning of the DP parameters, other running parameters can be tuned to improve reconstruction loss. Some of these relate to the underlying structure of the VAEs – namely the size of the vector  $Z$  of latent encoding, or the size of the hidden layer in the Neural Network. As seen in Figure 4.9, panels (b) and (c), while too small latent representation in  $Z$  may result in greater loss due to the inability to encode the data well enough, a value of  $Z$  which is too big may result in over-fitting, and might also result in higher reconstruction loss. Hidden layer dimensions, though, may allow for richer representation, but come at a cost of higher running times.

No noise											
$\epsilon$	$\delta$										
100	$10^{-2}$										
100	$10^{-5}$										
10	$10^{-2}$										
1	$10^{-2}$										
10	$10^{-5}$										
1	$10^{-5}$										

more noise

Table 4.1: Results of data fusion – MNIST dataset: 10 MNIST digits, with varying levels of noise added. The left side of each pair is the original digit. The right side is the reconstructed digit after splitting and fusing the pixels: the middle 300 pixels are common across datasets, and the rest are unique for each of the two datasets. The upper row is with no added noise. The next rows are with varying levels of  $\epsilon$  and  $\delta$  values. Smaller  $\epsilon$  means the privacy budget is lower, therefore more noise is added to the DP-SGD, as explained in Subsection 4.3.3.1.

## 4.5 Proposed Application: Anonymous Survey and CRM Data

To illustrate the managerial relevance of PPDF, we’ll now describe the planned fusion of two unique datasets, both stemming from a matchmaking website intended for extramarital affairs.

The data that will be used for estimation is from a website that specializes in extramarital relationships (hereafter “service” or “website”), a social network marketed for people seeking affairs. The website offers search tools for potential affair mates and several types of messaging tools. The goal of this exercise is to complement a rich behavioral CRM dataset with anonymous survey responses. The survey includes attitudes and stated preferences, and our primary goal with PPDF is to fuse them, as well as learn both about their joint distribution (e.g., learn how stated affair goals end up in eventual choices and attitudes towards affair-seeking (among those that actively engaged in it) without ever identifying a single user within either dataset.

The study was conducted in collaboration with the affair-seeking website to learn

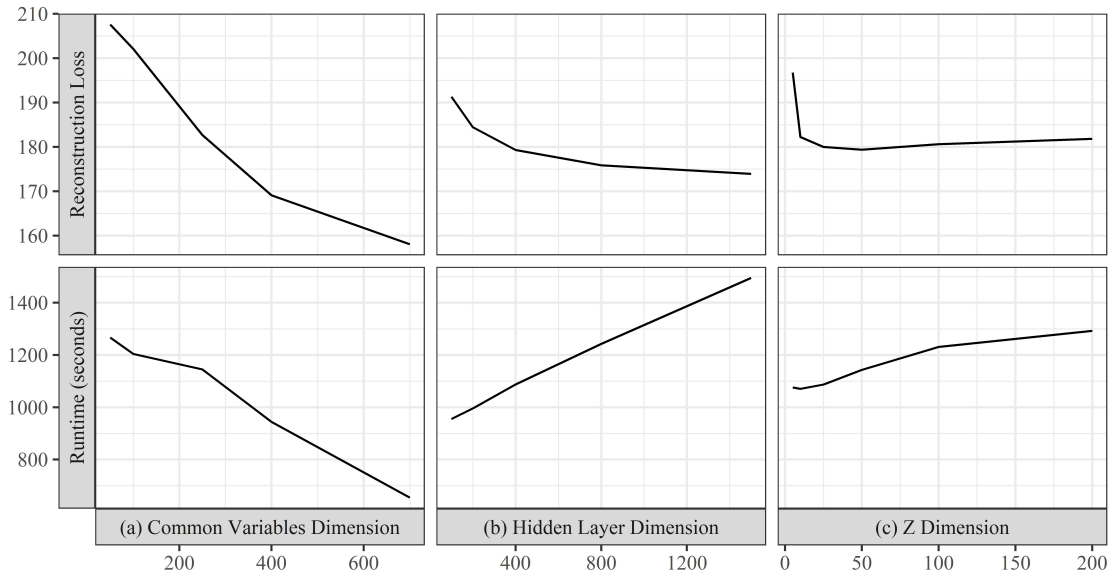


Figure 4.9: Simulation results – varying tuning parameters: reconstruction loss (the loss relative to the original images, upper panels) and running time (lower panels) as a function of the number of common variables (panel (a)), length of hidden layer (panel (b)) and length of  $Z$  – the latent representation of the encoders. Learning rate is  $\eta = 2 \cdot 10^{-4}$ . All simulations ran for 25 epochs. Number of common variables, if doesn't vary, is 300 pixels of the 784 pixels in each digit. Hidden Layer dimension, if doesn't vary, is 400. Latent representation –  $Z$  dimension – if doesn't vary, is 25.

more about attitudes and antecedents of online affair-seekers. Our research team<sup>6</sup>, similarly to many firms with a CRM system and a complementary dataset, aimed to apply data fusion techniques. Our goal was to fuse the detailed survey results with the rich behavioral CRM data, in order to gain greater insight on attitudes and behaviors of online affair-seekers. However, we soon recognized that customers' identities might effectively be uncovered (e.g., customer IDs, demographics, preferences). Though this is an extreme example of data sensitivity, where an actual risk of being revealed as an affair-seeker may result in personal and social backlash, many other data fusion exercises may result in identification of customers. Such revealed sensitive information may harm individuals, damage brand equity, as well as result in litigious

<sup>6</sup>Research team of overall project: Longxiu Tian, Dana Turjeman, Fred Feinberg, Elizabeth Bruch, and Dan Ariely. The proposed Privacy Preserving Data Fusion method is authored by Longxiu Tian and Dana Turjeman.

and regulatory ramifications. The proposed method will allow fusing of the survey with behavioral data, without harming the privacy of the customers.

We'll now describe the available data that will be used to illustrate PPDF's abilities.

#### 4.5.1 CRM Data

Male users of the service are either "Guests" or "Full Members". Both Guests and Full Members must register in order to enter the website, by providing valid email address and personal information, as will later be described. The website is so-called "Freemium" website; guests have minimal access to the service's features: they can use the search tool and explore the site, but cannot initiate any conversation. Full Members have, in addition to the search tools, full access to the site's features. In order to become a Full Member, male users need to pay for credits. They can use the credits to interact with people on the site, e.g., initiate conversations and/or send gifts. Women, on the other hand, do not need to pay in order to use these features, and are considered Full Members automatically. This strategy is used in order to encourage women to be active on the website.

A person who wishes to join the website must register with valid email address, and to fill out a profile, which contains:

1. Demographics such as country, state, city, zip, gender, ethnicity and date of birth.
2. Description of looks: eye color, hair color, weight, height, etc.
3. Sexual preferences, and preferences for type of affair/relationship that is being sought.

The data stem from all full members from the United States, that is, paid men and all women, who are registered as United States members. Therefore, our dataset

adequately captures all the website’s users from the U.S. who can send messages to one another.

The data were collected from January, 2014 through August, 2015. However, for evaluation purposes of this model, we use only the data of users that joined the website from January 1st, 2014 to October 31st, 2014, in their activities up to December 31st, approximately two months after the survey. This, in order to include users that were members of the website when the survey took place. In this pre-stated subsample, there are approximately one million users, all uniquely identified by their user id, a unique number that is assigned to a user upon registration and cannot be changed.

Some users altered some of their attributes (such as marital status, appearance, etc.) at different times after joining the website. Therefore, we decided to take the first observation of each user. If there is a missing value in this observation<sup>7</sup>, for the purposes of this exercise, we will complete the missing value from a later observation of the same user. The reason for taking the first observation is that it seems that some of the fields, such as date of birth, are being updated to a value that will be more attractive to other members, after joining the website. While the reasoning behind such an update will be discussed in a different scope, we hereby explicitly assume that the first input for each field is the one to better reflect the true attributes of the user.

Among all users in the database, women are 39% and men are 61%. 70.1% of users are Caucasian, 8.1% are Afro-American, 7.9% are Hispanic, 2.8% are Asian and 0.2% are First Nations (Native Americans); this attribute also allows for responses of “Other” or “Rather not say”. Figure 1 shows the distribution of ethnicity, compared to the survey respondents that will be described shortly.

---

<sup>7</sup>Users must fill in every field, other than optional text fields. Despite this obligation, the CRM data contain observations that have missing values. This is due to changes in the website — users that joined before certain time were allowed to leave some variables blank, and some data that we received from the company might have data that was recorded only when the user was first active, and not immediately upon joining. For the current illustration of PPDF, the missingness that is of most relevance is that of date of birth, where missing values account for less than 3%.

Our data contain individual level occurrences of messages, with both sender and receiver user IDs, as well as searches made on the website. The searches are especially rich, and may enable us to learn the preference of the users on the website, and the result: whether to message or not.

#### **4.5.2 Survey Responses**

We collaborated with the website, and conducted a unique survey among the website's users. All users on the websites received an email, and 5,461 users completed it within two days in 2014. Descriptions of the data are provided in a limited form, according to the restrictions of our Non-Disclosure Agreement.

During September and October 2014, we sent a survey link via an email to all registered United States users of the website, regardless of their membership status. The incentive for participation in the survey was that each participant who completed the survey is entered into a drawing to win 1,000 credits on the website, which is roughly equivalent to \$290. A limitation of the incentive is that women, who do not need credits in order to initiate contact on the website, are differentially incentivized to participate. This might explain why, among all users who completed the survey and filled in their gender, 95% are men and only 5% are women, even though the email was sent to all users.

Participants in the survey had the option to skip any questions they didn't feel comfortable responding to. None of the questions were mandatory, but, as can be seen in Appendix 4.7.1 and Appendix 4.7.2, survey respondents answered most of the questions. Questions that are relevant to subgroup of respondents, such as those that relate to spouses, have lower response rates because only participants who stated they have spouse were asked to answer such questions. Open ended questions received lower rate of responses (2%-14%) for obvious reasons.

## 4.6 Summary

In this paper, we presented a Privacy Preserving methodology for data fusion. Using a combination of state-of-the-art methodologies, we created a latent representation of the joint distribution of two completely remote datasets, with differential privacy built into the creation. Once constructed, any query regarding this joint distribution can be responded to.

The challenges of marketing automation and analytics in the era of data privacy are ongoing and multifaceted. This project aims to understand how data fusion, a prevalent marketing analytics technique, can be better retooled to meet today’s new privacy standards and practices. Our methodology offers a practical solution to collecting and storing less data. We show that collecting less data does not mean forgoing the advantages and insights that existing data fusion techniques allow. Using PPDF, companies can safely fuse datasets without their “ever meeting one another”, and potentially even split data, thus protecting customers’ fundamental right to privacy and reducing the risks associated with data breaches and leaks.

## 4.7 Appendices

### 4.7.1 A1: Summary Statistics of Survey

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
Accumulated Time (minutes) [numeric]	Mean (sd) : 18.6 (55.7) min < med < max: 0 < 10 < 1299 IQR (CV) : 6 (3)	178 distinct values		5530 (100.0%)	0 (0.0%)
Consent [factor]	1. No 2. Yes	71 ( 1.3%) 5459 (98.7%)		5530 (100.0%)	0 (0.0%)
Gender [factor]	1. Female 2. Male	267 ( 4.9%) 5157 (95.1%)		5424 (98.1%)	106 (1.9%)
Age group [factor]	1. Less than 20 2. 20-24 3. 25-29 4. 30-34 5. 35-39 6. 40-44 7. 45-49 8. 50-54 9. 55-59 10. 60-64 [ 2 others ]	3 ( 0.1%) 47 ( 0.9%) 151 ( 2.8%) 409 ( 7.5%) 617 (11.3%) 860 (15.8%) 923 (17.0%) 920 (16.9%) 701 (12.9%) 467 ( 8.6%) 345 ( 6.3%)		5443 (98.4%)	87 (1.6%)

Figure 4.10: Survey Summary Table #1



Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
Race [factor]	1. African American 2. Asian 3. Hispanic 4. Native American 5. Other 6. Pacific Islander 7. White/Caucasian	259 ( 4.8%) 83 ( 1.5%) 398 ( 7.3%) 35 ( 0.6%) 137 ( 2.5%) 21 ( 0.4%) 4500 (82.8%)		5433 (98.2%)	97 (1.8%)
Education [factor]	1. Some high school 2. High school graduate 3. Some college 4. College graduate 5. Postgraduate/professional	51 ( 0.9%) 500 ( 9.2%) 1630 (30.0%) 2036 (37.5%) 1213 (22.3%)		5430 (98.2%)	100 (1.8%)
Work_status [factor]	1. No 2. Yes, full time 3. Yes, part time	598 (11.0%) 4489 (82.9%) 328 ( 6.1%)		5415 (97.9%)	115 (2.1%)
HH income [factor]	1. Under \$25,000 2. \$25,001 - \$49,999 3. \$50,000 - \$74,999 4. \$75,000 - \$99,999 5. \$100,000 - \$149,999 6. \$150,000 - \$200,000 7. \$200,000 - \$250,000 8. More than \$250,000	208 ( 3.9%) 721 (13.4%) 1027 (19.1%) 1094 (20.3%) 1286 (23.9%) 555 (10.3%) 191 ( 3.5%) 299 ( 5.6%)		5381 (97.3%)	149 (2.7%)

Figure 4.11: Survey Summary Table #2

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
Political [factor]	1. Democrat	1247 (23.1%)		5409 (97.8%)	121 (2.2%)
	2. Independent	1936 (35.8%)			
	3. Other [specify]	297 ( 5.5%)			
	4. Republican	1929 (35.7%)			
Religion [factor]	1. Atheist	623 (11.5%)		5397 (97.6%)	133 (2.4%)
	2. Buddhist	74 ( 1.4%)			
	3. Catholic	1488 (27.6%)			
	4. Hindu	25 ( 0.5%)			
	5. Jewish	176 ( 3.3%)			
	6. Muslim/Islam	22 ( 0.4%)			
	7. Other	1527 (28.3%)			
	8. Protestant	1462 (27.1%)			
How often religious services [factor]	1. Never	2124 (39.2%)		5422 (98.0%)	108 (2.0%)
	2. Less than Once a Month	1838 (33.9%)			
	3. Once a Month	394 ( 7.3%)			
	4. 2-3 Times a Month	457 ( 8.4%)			
	5. Once a Week	471 ( 8.7%)			
	6. 2-3 Times a Week	100 ( 1.8%)			
	7. Daily	38 ( 0.7%)			

Figure 4.12: Survey Summary Table #3

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
Relationship status [factor]	1. Casual relationship 2. Living with a partner 3. Married 4. Open 5. Other 6. Seperated 7. Serious relationship 8. Single [never married, di	244 ( 4.5%) 392 ( 7.2%) 3364 (61.9%) 4 ( 0.1%) 83 ( 1.5%) 41 ( 0.8%) 191 ( 3.5%) 1112 (20.5%)		5431 (98.2%)	99 (1.8%)
Ever married [factor]	1. No 2. Yes	602 (29.0%) 1471 (71.0%)		2073 (37.5%)	3457 (62.5%)
First time marriage age [factor]	1. less than 20 2. 20-25 3. 25-30 4. 30-35 5. 35-40 6. older than 40	453 ( 9.4%) 2124 (44.1%) 1344 (27.9%) 664 (13.8%) 161 ( 3.3%) 66 ( 1.4%)		4812 (87.0%)	718 (13.0%)
Times married [numeric]	Mean (sd) : 1.2 (0.7) min < med < max: 0 < 1 < 3 IQR (CV) : 1 (0.6)	0 : 602 (11.1%) 1 : 3382 (62.4%) 2 : 1113 (20.5%) 3 : 326 ( 6.0%)		5423 (98.1%)	107 (1.9%)

Figure 4.13: Survey Summary Table #4

#### 4.7.2 A2: Response Rates for a Sample of Survey Questions

Question	Responses	Percent
What is your gender?	5424	99%
How old are you?	5443	100%
What is your race?	5433	99%
What is your race?-TEXT	84	2%
What is your highest level of education?	5430	99%
Are you currently working for pay?	5415	99%
Please indicate your approximate yearly household income before taxes	5381	99%
Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or some...	5409	99%
Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or some...-TEXT	260	5%
Which of the following best represents how you think about your own religion or spiritual life?	5397	99%
Which of the following best represents how you think about your own religion or spiritual life? -TEXT	756	14%
To what extent do you consider yourself a religious person?	4801	88%
How often do you take part in the services, activities, and social life of a church or place of worship?	5422	99%
What is your current relationship status?	5431	99%
What is your current relationship status? -TEXT	115	2%
Have you ever been married?	2073	38%
How many times have you been married?	1468	27%
Including your present marriage, how many times have you been married?	3353	61%
When you married for the first time, how old were you?	4812	88%
Does your current partner/spouse know you are on the site?	3931	72%
How long have you been with your current partner/spouse?	3933	72%
What is the highest level of education that your husband/wife/partner completed?	3928	72%
How old is your husband/wife/partner?	3930	72%
What race or ethnicity is your husband/wife/partner?	3929	72%

## BIBLIOGRAPHY

- Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang (2016), Deep learning with differential privacy, in Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp. 308–318.
- Ablon, L., P. Heaton, D. C. Lavery, and S. Romanosky (2016), Consumer attitudes toward data breach notifications and loss of personal information, Rand Corporation.
- Acquisti, A., and H. R. Varian (2005), Conditioning prices on purchase history, *Marketing Science*, 24(3), 367–381.
- Acquisti, A., A. Friedman, and R. Telang (2006), Is there a cost to privacy breaches? an event study, *ICIS 2006 Proceedings*, p. 94.
- Acquisti, A., S. Gritzalis, C. Lambrinoudakis, and S. di Vimercati (2007), *Digital privacy: theory, technologies, and practices*, CRC Press.
- Acquisti, A., et al. (2017), Nudges for privacy and security: Understanding and assisting users’ choices online, *ACM Computing Surveys (CSUR)*, 50(3), 1–41.
- Almuhimedi, H., F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal (2015), Your location has been shared 5,398 times! a field study on mobile app privacy nudging, in Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 787–796.
- Amir, E., S. Levi, and T. Livne (2018), Do firms underreport information on cyberattacks? evidence from capital markets, *Review of Accounting Studies*, 23(3), 1177–1206.
- Athey, S., and S. Wager (2019), Estimating treatment effects with causal forests: An application, *arXiv preprint arXiv:1902.07409*.
- Athey, S., C. Catalini, and C. Tucker (2017), The digital privacy paradox: Small money, small costs, small talk, Tech. rep., National Bureau of Economic Research, working paper series.
- Athey, S., J. Tibshirani, S. Wager, et al. (2019), Generalized random forests, *Annals of Statistics*, 47(2), 1148–1178.

- Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021), Matrix completion methods for causal panel data models, *Journal of the American Statistical Association*, pp. 1–41.
- Bagdasaryan, E., A. Veit, Y. Hua, D. Estrin, and V. Shmatikov (2020), How to backdoor federated learning, in *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948, PMLR.
- Bai, J. (2009), Panel data models with interactive fixed effects, *Econometrica*, 77(4), 1229–1279.
- Bakken, D. E., R. Rameswaran, D. M. Blough, A. A. Franz, and T. J. Palmer (2004), Data obfuscation: Anonymity and desensitization of usable data sets, *IEEE Security & Privacy*, 2(6), 34–41.
- Berry, S., J. Levinsohn, and A. Pakes (2004), Differentiated products demand systems from a combination of micro and macro data: The new car market, *Journal of political Economy*, 112(1), 68–105.
- Bertrand, M., E. Duflo, and S. Mullainathan (2004), How much should we trust differences-in-differences estimates?, *The Quarterly journal of economics*, 119(1), 249–275.
- Bogost, I. (2018), Welcome to the age of privacy nihilism, *The Atlantic*.
- Bonawitz, K., V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth (2017), Practical secure aggregation for privacy-preserving machine learning, in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191.
- Bradburn, N. M., S. Sudman, E. Blair, W. Locander, C. Miles, E. Singer, and C. Stocking (1979), Improving interview method and questionnaire design: Response effects to threatening questions in survey research, *Jossey-Bass San Francisco*.
- Bradlow, E. T., and A. M. Zaslavsky (1999), A hierarchical latent variable model for ordinal data from a customer satisfaction survey with “no answer” responses, *Journal of the American Statistical Association*, 94(445), 43–52.
- Breiman, L. (2001), Random forests, *Machine learning*, 45(1), 5–32.
- Brisson, S. (2018), Canada’s national action plan on open government, *The Government of Canada*.
- Bruch, E. E., and M. Newman (2018), Aspirational pursuit of mates in online dating markets, *Science Advances*, 4(8), eaap9815.
- Buolamwini, J., and T. Gebru (2018), Gender shades: Intersectional accuracy disparities in commercial gender classification, in *Conference on fairness, accountability and transparency*, pp. 77–91, PMLR.

- Cadwalladr, C. (2017), The great british brexit robbery: how our democracy was hijacked, *The Guardian*, 20.
- Caspit, B. (2018), Cyber attacks increasingly penetrating israel's politics, *Al Monitor*.
- Chandar, S., M. M. Khapra, H. Larochelle, and B. Ravindran (2016), Correlational neural networks, *Neural computation*, 28(2), 257–285.
- Choong, P., E. Hutton, P. Richardson, and V. Rinaldo (2016), Assessing the cost of security breach: A marketer's perspective, in *Allied Academies International Conference. Academy of Marketing Studies. Proceedings*, vol. 21, p. 1, Jordan Whitney Enterprises, Inc.
- Cleeren, K., M. G. Dekimpe, and H. J. van Heerde (2017), Marketing research on product-harm crises: a review, managerial implications, and an agenda for future research, *Journal of the Academy of Marketing Science*, 45(5), 593–615.
- Confessore, N. (2018), Cambridge analytica and facebook: The scandal and the fallout so far, *New York Times*.
- Cowgill, B., and C. E. Tucker (2019), Economics, fairness and algorithmic bias, preparation for: *Journal of Economic Perspectives*.
- Cox, D. R. (1958), *Planning of experiments*, New York.
- Dance, G. J. X., M. LaForgia, and N. Confessore (2018), As facebook raised a privacy wall, it carved an opening for tech giants, *New York Times*.
- Dautov, R., S. Distefano, and R. Buyya (2019), Hierarchical data fusion for smart healthcare, *Journal of Big Data*, 6(1), 1–23.
- Dellaert, B. G., et al. (2018), Individuals' decisions in the presence of multiple goals, *Customer Needs and Solutions*, 5(1), 51–64.
- Deng, L. (2012), The mnist database of handwritten digit images for machine learning research [best of the web], *IEEE Signal Processing Magazine*, 29(6), 141–142.
- Dias, F. F., P. S. Lavieri, T. Kim, C. R. Bhat, and R. M. Pendyala (2019), Fusing multiple sources of data to understand ride-hailing use, *Transportation Research Record*, 2673(6), 214–224.
- DiResta, R., K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, and B. Johnson (2019), The tactics & tropes of the internet research agency, *New Knowledge*.
- Domingo-Ferrer, J., and V. Torra (2008), A critique of k-anonymity and some of its enhancements, in *2008 Third International Conference on Availability, Reliability and Security*, pp. 990–993, IEEE.

- Dooley, S., T. Goldstein, and J. P. Dickerson (2021), Robustness disparities in commercial face detection, Working paper.
- Dunn, H. L. (1946), Record linkage, *American Journal of Public Health and the Nations Health*, 36(12), 1412–1416.
- Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor (2006a), Our data, ourselves: Privacy via distributed noise generation, in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, Springer.
- Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006b), Calibrating noise to sensitivity in private data analysis, in *Theory of cryptography conference*, pp. 265–284, Springer.
- eMarketer (2014), Retailers need to beef up security to regain trust.
- Evans, G., G. King, M. Schwenzfeier, and A. Thakurta (2020), Statistically valid inferences from privacy protected data, URL: [GaryKing.org/dp](http://GaryKing.org/dp).
- Feit, E. M., M. A. Beltramo, and F. M. Feinberg (2010), Reality check: Combining choice experiments with market data to estimate the importance of product attributes, *Management science*, 56(5), 785–800.
- Feit, E. M., P. Wang, E. T. Bradlow, and P. S. Fader (2013), Fusing aggregate and disaggregate data with an application to multiplatform media consumption, *Journal of Marketing Research*, 50(3), 348–364.
- Foeman, A., B. L. Lawton, and R. Rieger (2015), Questioning race: Ancestry dna and dialog on race, *Communication Monographs*, 82(2), 271–290.
- Friedberg, R., J. Tibshirani, S. Athey, and S. Wager (2020), Local linear forests, *Journal of Computational and Graphical Statistics*, pp. 1–15.
- Fung, B. C., K. Wang, R. Chen, and P. S. Yu (2010), Privacy-preserving data publishing: A survey of recent developments, *ACM Computing Surveys (Csur)*, 42(4), 1–53.
- Gilula, Z., R. E. McCulloch, and P. E. Rossi (2006), A direct approach to data fusion, *Journal of Marketing Research*, 43(1), 73–83.
- Goldfarb, A., and C. Tucker (2011), Online display advertising: Targeting and obtrusiveness, *Marketing Science*, 30(3), 389–404.
- Goldfarb, A., and C. Tucker (2012), Shifts in privacy concerns, *American Economic Review*, 102(3), 349–53.
- Golle, P. (2006), Revisiting the uniqueness of simple demographics in the us population, in *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, pp. 77–80.



- Gordon, L. A., M. P. Loeb, and L. Zhou (2011), The impact of information security breaches: Has there been a downward shift in costs?, *Journal of Computer Security*, 19(1), 33–56.
- Graham, J. W. (2009), Missing data analysis: Making it work in the real world, *Annual review of psychology*, 60, 549–576.
- Granger, C. W. (1980), Testing for causality: a personal viewpoint, *Journal of Economic Dynamics and control*, 2, 329–352.
- Guadagni, P. M., and J. D. Little (1983), A logit model of brand choice calibrated on scanner data, *Marketing science*, 2(3), 203–238.
- Guo, T., S. Sriram, and P. Manchanda (2021), The effect of information disclosure on industry payments to physicians., *Journal of Marketing Research (JMR)*, 58(1).
- Han, J. A., E. M. Feit, and S. Srinivasan (2020), Can negative buzz increase awareness and purchase intent?, *Marketing Letters*, 31(1), 89–104.
- Harrell, E., B. of Justice Statistics (BJS), U. D. of Justice, O. of Justice Programs, and U. S. of America (2015), Victims of identity theft, 2014, Bureau of Justice Statistics (BJS).
- Janakiraman, R., J. H. Lim, and R. Rishika (2018), The effect of a data breach announcement on customer behavior: Evidence from a multichannel retailer, *Journal of Marketing*, 82(2), 85–105.
- Johnson, E. J., et al. (2012), Beyond nudges: Tools of a choice architecture, *Marketing Letters*, 23(2), 487–504.
- Kamakura, W. A., and M. Wedel (1997), Statistical data fusion for cross-tabulation, *Journal of Marketing Research*, 34(4), 485–498.
- Kamakura, W. A., and M. Wedel (2000), Factor analysis and missing data, *Journal of Marketing Research*, 37(4), 490–498.
- Kim, S., C. Lee, and S. Gupta (2020), Bayesian synthetic control methods, *Journal of Marketing Research*, 57(5), 831–852.
- Kim, T., K. Barasz, and L. K. John (2019), Why am i seeing this ad? the effect of ad transparency on ad effectiveness, *Journal of Consumer Research*, 45(5), 906–932.
- Kingma, D. P., and M. Welling (2019), An introduction to variational autoencoders, *arXiv preprint arXiv:1906.02691*.
- Kiritchenko, S., and S. M. Mohammad (2018), Examining gender and race bias in two hundred sentiment analysis systems, *arXiv preprint arXiv:1805.04508*.
- Klami, A., S. Virtanen, and S. Kaski (2013), Bayesian canonical correlation analysis., *Journal of Machine Learning Research*, 14(4).

- Knaus, M. C., M. Lechner, and A. Strittmatter (2021), Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence, *The Econometrics Journal*, 24(1), 134–161.
- Kude, T., H. Hoehle, and T. A. Sykes (2017), Big data breaches and customer compensation strategies: Personality traits and social influence as antecedents of perceived compensation, *International Journal of Operations & Production Management*.
- Läll, K., R. Mägi, A. Morris, A. Metspalu, and K. Fischer (2017), Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores, *Genetics in Medicine*, 19(3), 322–329.
- Lambrecht, A., and C. Tucker (2021), Algorithm-based advertising: Unintended effects and the tricky business of mitigating adverse outcomes., *NIM Marketing Intelligence Review*, 13(1).
- Langheinrich, M., and F. Schaub (2018), Privacy in mobile and pervasive computing, *Synthesis Lectures on Mobile and Pervasive Computing*, 10(1), 1–139.
- Lee, C., and P. Anand (2020), Using deep learning to overcome privacy and scalability issues in customer data transfer, Available at SSRN 3769521.
- Li, L., B. Pal, J. Ali, N. Sullivan, R. Chatterjee, and T. Ristenpart (2019), Protocols for checking compromised credentials, in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1387–1403.
- Li, N., T. Li, and S. Venkatasubramanian (2007), t-closeness: Privacy beyond k-anonymity and l-diversity, in *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115, IEEE.
- Liang, F., V. Das, N. Kostyuk, and M. M. Hussain (2018), Constructing a data-driven society: China’s social credit system as a state surveillance infrastructure, *Policy & Internet*, 10(4), 415–453.
- Lin, C., D. He, X. Huang, M. K. Khan, and K.-K. R. Choo (2018), A new transitively closed undirected graph authentication scheme for blockchain-based identity management systems, *IEEE Access*, 6, 28,203–28,212.
- Little, R. J., and D. B. Rubin (1989), The analysis of social science data with missing values, *Sociological Methods & Research*, 18(2-3), 292–326.
- Liu, J., T. Li, P. Xie, S. Du, F. Teng, and X. Yang (2020), Urban big data fusion based on deep learning: An overview, *Information Fusion*, 53, 123–133.
- Machanavajjhala, A., D. Kifer, J. Gehrke, and M. Venkatasubramanian (2007), l-diversity: Privacy beyond k-anonymity, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 3–es.

- Madden, M., and L. Rainie (2015), Americans' attitudes about privacy, security and surveillance, Pew Research Center.
- Mansfield-Devine, S. (2016), The secure way to use open source, *Computer Fraud & Security*, 2016(5), 15–20.
- Marino, S., N. Zhou, Y. Zhao, L. Wang, Q. Wu, and I. D. Dinov (2019), Hdda: Datasifter: statistical obfuscation of electronic health records and other sensitive datasets, *Journal of statistical computation and simulation*, 89(2), 249–271.
- Massey Jr, F. J. (1951), The kolmogorov-smirnov test for goodness of fit, *Journal of the American statistical Association*, 46(253), 68–78.
- McCarthy, D. M., and E. S. Oblander (2021), Scalable data fusion with selection correction: An application to customer base analysis, *Marketing Science*.
- McDonald, A. M., and L. F. Cranor (2008), The cost of reading privacy policies, *Isjlp*, 4, 543.
- McNair, C. (2018), Global digital users update, eMarketer, 2018.
- Miller, A. R. (1969), Personal privacy in the computer age: The challenge of a new technology in an information-oriented society, *Michigan Law Review*, 67(6), 1089–1246.
- Molitor, D., M. Spann, A. Ghose, and P. Reichhart (2020), Effectiveness of location-based advertising and the impact of interface design, *Journal of Management Information Systems*, 37(2), 431–456.
- Nahum-Shani, I., A. Ertefaie, X. Lu, K. G. Lynch, J. R. McKay, D. W. Oslin, and D. Almirall (2017), A smart data analysis method for constructing adaptive treatment strategies for substance use disorders, *Addiction*, 112(5), 901–909.
- Nalls, M. A., et al. (2014), Large-scale meta-analysis of genome-wide association data identifies six new risk loci for parkinson's disease, *Nature genetics*, 46(9), 989–993.
- Narayanan, A., and V. Shmatikov (2008), Robust de-anonymization of large sparse datasets, in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, IEEE.
- Nissenbaum, H. (2009), *Privacy in context*, Stanford University Press.
- Norberg, P. A., D. R. Horne, and D. A. Horne (2007), The privacy paradox: Personal information disclosure intentions versus behaviors, *Journal of consumer affairs*, 41(1), 100–126.
- Obar, J. A., and A. Oeldorf-Hirsch (2020), The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services, *Information, Communication & Society*, 23(1), 128–147.

- of State Legislatures, N. C. (2018), Security breaches notification laws, National Conference of State Legislatures.
- Papamakarios, G., E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan (2019), Normalizing flows for probabilistic modeling and inference, arXiv preprint arXiv:1912.02762.
- Papernot, N. (2019), Machine learning at scale with differential privacy in tensorflow, in 2019 {USENIX} Conference on Privacy Engineering Practice and Respect ({PEPR} 19), pp. 1–1.
- Ping, H., J. Stoyanovich, and B. Howe (2017), Datasynthesizer: Privacy-preserving synthetic datasets, in Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 1–5.
- Posner, R. A. (1981), The economics of privacy, *The American economic review*, 71(2), 405–409.
- Qian, Y., and H. Xie (2014), Which brand purchasers are lost to counterfeiters? an application of new data fusion approaches, *Marketing Science*, 33(3), 437–448.
- Raji, I. D., and J. Buolamwini (2019), Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products, in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 429–435.
- Ram, N., C. J. Guerrini, and A. L. McGuire (2018), Genealogy databases and the future of criminal investigation, *Science*, 360(6393), 1078–1079.
- Rezende, D., and S. Mohamed (2015), Variational inference with normalizing flows, in International Conference on Machine Learning, pp. 1530–1538, PMLR.
- Rezende, D. J., and F. Viola (2018), Taming vaes, arXiv preprint arXiv:1810.00597.
- Romanosky, S., R. Telang, and A. Acquisti (2011), Do data breach disclosure laws reduce identity theft?, *Journal of Policy Analysis and Management*, 30(2), 256–286.
- Romanosky, S., D. Hoffman, and A. Acquisti (2014), Empirical analysis of data breach litigation, *Journal of Empirical Legal Studies*, 11(1), 74–104.
- Rosati, P., M. Cummins, P. Deeney, F. Gogolin, L. van der Werff, and T. Lynn (2017), The effect of data breach announcements beyond the stock price: Empirical evidence on market activity, *International Review of Financial Analysis*, 49, 146–154.
- Rosenfeld, M. J., and R. J. Thomas (2012), Searching for a mate: The rise of the internet as a social intermediary, *American Sociological Review*, 77(4), 523–547.
- Rubin, D. B. (1976), Inference and missing data, *Biometrika*, 63(3), 581–592.

- Rubin, D. B. (1980), Randomization analysis of experimental data: The fisher randomization test comment, *Journal of the American Statistical Association*, 75(371), 591–593.
- Rubin, D. B. (1990), Formal mode of statistical inference for causal effects, *Journal of statistical planning and inference*, 25(3), 279–292.
- Rubin, D. B. (2005), Causal inference using potential outcomes: Design, modeling, decisions, *Journal of the American Statistical Association*, 100(469), 322–331, doi: 10.1198/016214504000001880.
- Ryffel, T., A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach (2018), A generic framework for privacy preserving deep learning, arXiv preprint arXiv:1811.04017.
- Schneider, M. J., S. Jagpal, S. Gupta, S. Li, and Y. Yu (2017), Protecting customer privacy when marketing with second-party data, *International Journal of Research in Marketing*, 34(3), 593–603.
- Schneider, M. J., S. Jagpal, S. Gupta, S. Li, and Y. Yu (2018), A flexible method for protecting marketing data: An application to point-of-sale data, *Marketing Science*, 37(1), 153–171.
- Scott, T., and A. Rung (2016), Memorandum for heads of departments and agencies: Federal source code policy: Achieving efficiency, transparency, and innovation through reusable and open source software, U.S. General Services Administration, pp. 16–21.
- Simmons, R. (2017), Big data and procedural justice: Legitimizing algorithms in the criminal justice system, *Ohio St. J. Crim. L.*, 15, 573.
- Sprenger, P. (1999), Sun on privacy: ‘get over it’, *Wired News*, 26, 1–4.
- Starr, D. (2016), When dna is lying.
- Stigler, G. J. (1980), An introduction to privacy in economics and politics, *The Journal of Legal Studies*, 9(4), 623–644.
- Swait, J., and R. L. Andrews (2003), Enriching scanner panel models with choice experiments, *Marketing Science*, 22(4), 442–460.
- Sweeney, L. (1997), Weaving technology and policy together to maintain confidentiality, *The Journal of Law, Medicine & Ethics*, 25(2-3), 98–110.
- Sweeney, L. (2000), Simple demographics often identify people uniquely, *Health (San Francisco)*, 671(2000), 1–34.
- Sweeney, L. (2002), k-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05), 557–570.

- Synoposys (2015), Coverity scan open source report shows commercial code is more compliant to security standards than open source code, Synopsis.
- Takagi, S., T. Takahashi, Y. Cao, and M. Yoshikawa (2020), P3gm: Private high-dimensional data release via privacy preserving phased generative model, arXiv preprint arXiv:2006.12101.
- Taylor, C. R. (2004), Consumer privacy and the market for customer information, *RANDJ. Econom.*, 35(4), 631–651.
- The European Union (2016-05-04), Regulation (eu) 2016/679 of the european parliament and of the council 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, and repealing directive 95/46/ec (general data protection regulation), OJ, L 119, 1.
- Tucker, C. E. (2014), Social networks, personalized advertising, and privacy controls, *Journal of marketing research*, 51(5), 546–562.
- Turjeman, D., and F. M. Feinberg (2021), When the data are out: measuring behavioral changes following a data breach, working paper.
- Turjeman, D., and L. Tian (2021), Privacy Preserving Data Fusion, working paper.
- Wager, S., and S. Athey (2018), Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association*.
- Wager, S., T. Hastie, and B. Efron (2014), Confidence intervals for random forests: The jackknife and the infinitesimal jackknife, *The Journal of Machine Learning Research*, 15(1), 1625–1651.
- Westin, A. F. (1968), Privacy and freedom, *Washington and Lee Law Review*, 25(1), 166.
- Wurmser, Y. (2018), Maps and navigation apps – discovery, exploration features open up ad opportunities, eMarketer, 2018.
- Xu, Y. (2017), Generalized synthetic control method: Causal inference with interactive fixed effects models, *Political Analysis*, 25(1), 57–76.
- Zeevi, D., et al. (2015), Personalized nutrition by prediction of glycemic responses, *Cell*, 163(5), 1079–1094.
- Zheng, Y. (2015), Methodologies for cross-domain data fusion: An overview, *IEEE transactions on big data*, 1(1), 16–34.
- Zhong, N., and D. A. Schweidel (2020), Capturing changes in social media content: a multiplelatent changepoint topic model, *Marketing Science*, 39(4), 827–846.

Zou, Y., and F. Schaub (2018), Concern but no action: Consumers' reactions to the equifax data breach, in Extended abstracts of the 2018 CHI conference on human factors in computing systems, pp. 1–6.

Zou, Y., and F. Schaub (2019), Beyond mandatory: Making data breach notifications useful for consumers, *IEEE Security & Privacy*, 17(2), 67–72.

Zuboff, S. (2019), The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019, Profile books.