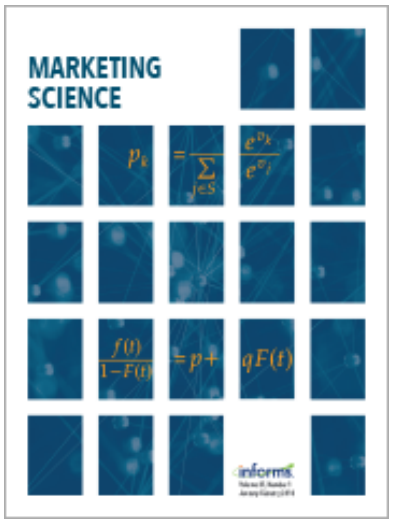


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Dana Turjeman,, Fred M. Feinberg

To cite this article:
Dana Turjeman,, Fred M. Feinberg (2023) When the Data Are Out: Measuring Behavioral Changes Following a Data Breach. Marketing Science

Published online in Articles in Advance 16 Aug 2023

. https://doi.org/10.1287/mksc.2019.0208

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–22
ISSN 0732-2399 (print), ISSN 1526-548X (online)

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

Dana Turjeman,[a,*] Fred M. Feinberg[b]

[a] Arison School of Business, Reichman University, Herzliya 4610101, Israel; [b] Ross School of Business and Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109

*Corresponding author

Contact: dana.turjeman@runi.ac.il, https://orcid.org/0000-0003-1445-2983 (DT); feinf@umich.edu, https://orcid.org/0000-0003-2238-0721 (FMF)

Received: July 17, 2019
Revised: July 21, 2022; January 21, 2023; April 23, 2023
Accepted: April 26, 2023
Published Online in Articles in Advance: August 16, 2023

https://doi.org/10.1287/mksc.2019.0208

Abstract. As the quantity and value of data increase, so do the severity of data breaches and customer privacy invasions. Although firms typically publicize their post hoc protective actions, little is known about the aftereffects of major breaches on users' behaviors; do they alter their interactions with the firm, continue "business as usual," or do something more subtle? We explore these questions in the context of a severe data breach to a matchmaking website for those seeking an (extramarital) affair. A challenge to measuring "treatment effects" for a massive and highly publicized breach is the lack of an obvious control group. To resolve this problem, we propose *Temporal Causal Inference* (TCI); each group of users who joined during a given time window is matched with an appropriate (control) group of users who had joined prior to it, helping to account for "usage trajectories" in both individual and temporal site behavior. Following the creation of the control groups, we adapt Causal Forests (Athey et al. 2019) into *Temporal Causal Forests* (TCF). TCF allows for insights regarding both average and individual-level treatment (data breach) effects as well as both demographic and usage-based covariates that align with them. Our analyses reveal a decrease in the probability of searching and messaging on the website and a notable increase in the probability of deleting photos, the primary avenue for avoiding further personal identification. Moreover, these effects are broadly robust to a variety of causal inference methodologies, both with and without TCI or Causal Forests. Intriguingly, these initially negative reaction(s) did not persist; by the third week after the announcement, there were hints of "life returns to normal." Despite the specificity of the setting, our analysis suggests both managerial and policy imperatives to help protect customers' privacy.

History: Avi Goldfarb and Olivier Toubia served as the senior editor for this article.
Supplemental Material: The e-companion and data are available at https://doi.org/10.1287/mksc.2019.0208.

Keywords: data breaches • privacy • exogenous shocks • usage trajectories • causal inference

1. Introduction

The number of customer records revealed to be compromised in data breaches has increased dramatically in recent years. In 2012, some 20 million records were compromised; in 2015 the number rose to 318 million; in 2016 through 2018 the number reached the billions. Between 2019 and 2022, hundreds of known, publicized breaches entailed enormous numbers of individual records.[1] Despite the severity of some of these incidents, 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 firms nor their customers can anticipate—disclosures of data breaches may result in heterogeneous reactions among users and customers.[2] 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, Gwebu et al. 2018, Chen and Jai 2021), 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, indeed, any exogenous shock—is critical for firms, customers who rely on their protection, and policymakers crafting guidelines to mitigate potential damage. However, such measures are challenging to enact; in highly publicized shocks, it is uncommon for a group of users to remain uninformed and unaffected and thereby serve as controls. Moreover, post-breach measures of users' behaviors typically remain the sole possession of the breached firm, which is naturally reluctant to reveal

them. This reluctance hinders the ability of policymakers and managers to assist customers in remedying data disclosures or to safeguard them from future ones; making such analyses public is a crucial step in reducing harm for both customers and the affected company.

A common assumption following a major data breach is that the firm would not be able to recover, having squandered the trust of current and potential customers. Although in some cases this might be so, we provide evidence that, even in a data breach as uncommonly severe as the one we present, there is already some degree of attenuation in the negative effects just three weeks after the announcement. Specifically, our data stem from a matchmaking website that suffered a data breach so severe as to generate worldwide media attention. This attention was owed in large part to its intended audience: people seeking extramarital affairs. That is, the breached, disclosed data exposed users' desire to engage in a clandestine relationship, along with personally identifying information for all paying members.[3]

Because the data breach was massive and highly publicized, and all customers were directly informed of it, there is no obvious control group against which to measure its causal effect(s) on individual users' engagement with the firm's site. To overcome this, we extend Causal Forests methodology (Wager and Athey 2018, Athey et al. 2019), with *Temporal Causal Inference* (TCI) identification, a combination henceforth denoted as Temporal Causal Forests (TCF), where users' projected engagement is based on those who joined before them via fully nonparametric "trajectory matching." Although our main results are based on this methodology for assessing heterogeneous treatment effects when all users were effectively treated, we also check their substantive robustness via a wide range of alternative methods: differences in differences, regression discontinuity in time, and generalized and Bayesian synthetic control, among others. We discuss the advantages of TCI and TCF above and beyond these methods, although they may prove useful in other domains.

Our results suggest that users were less likely to engage in searches and messages on the website immediately after the data breach was announced and were far more likely to delete their photos than they otherwise would, which might be expected, given that these were among the only avenues to protect themselves from further identification post-announcement. However, by only the third week after the breach announcement, all such effects began to attenuate. Although this might be unsurprising for photos—deleted photos remain deleted—for searches and messages, even though the strongest reactions were in the second week after the breach, by the third week there were hints of "returning to normal;" that is, although the number of activities was still lower than it would have been *sans* treatment, the magnitude of the effect was becoming smaller. Although we do not observe detailed customer data for more than three weeks, the website still exists to date and may have benefited from the increased awareness (among both existing and new users) to the existence of the brand, and we discuss such possibilities in the Conclusion.

We also find variability in users' reactions to the breach and explore potential correlates of this heterogeneity. For example, users who had higher preferences for privacy were the last to react and had weaker reactions relative to other users, even though their privacy preferences (not sharing photos unless others shared back) were protective against neither future information breaches nor "curious" users who may have joined the website following the breach.

The remainder of the paper is organized as follows. Sections 1.1 and 1.2 review the literature on the effects of data breaches from the perspectives of the individual and of the company, along with the challenges to assess such exogenous shocks in a company's perceptions. Section 2 describes our data, whereas Section 3.1 outlines the construction of control groups via *Temporal Causal Inference*. Section 3.2 describes *Temporal Causal Forests*, a nonparametric approach to assessing individual treatment effects using Causal Forests. Section 4 details the results of these analyses, Sections 5–8 provide various robustness checks, and Section 9 closes by discussing the results and their implications.

1.1. Literature Review: Data Breaches

Because future data breaches are inevitable, public notification of their occurrence has been regulated in many countries. Whereas evidence from laboratory studies has suggested that data breach notifications can be neither clear nor particularly alarming to those victimized by them (Zou and Schaub 2019), Romanosky et al. (2011) found that such notices successfully reduced the number of identity thefts caused by data breaches by 6.1%.

To the individual affected, however, data breaches may entail more than just financial losses. They are perceived as privacy invasions, leading to lack of trust and information leaks of many sorts: purchase behavior, daily routines, email correspondences, and other forms of communications that were meant to remain private. People differ in both their stated and empirically demonstrated preferences for sharing personal information (see, for example, Taylor 2004, Acquisti and Varian 2005, Goldfarb and Tucker 2012, Athey et al. 2017, and Lin 2022). Such heterogeneity appears not only in willingness to share information with a target firm but also in the reactions following notification of a breach. According to a survey by Ablon et al. (2016), 89% of respondents continued to conduct business with a breached firm, whereas only 11% stopped cold, and 1% reported increasing the amount of business they

conducted with the breached firm. Following Facebook's allegations of privacy misconduct in the infamous "Cambridge Analytica" case, roughly half of Facebook's American users said they had not changed their degree of engagement with the site, a quarter said they were using it more, and the remaining quarter claimed they were using it less, stopped using it, or even deleted their account.[4]

Following the 2012 data breach to Target, which affected 70 million customers, Kude et al. (2017) surveyed 212 whose data were compromised, finding substantial differences in the perception of compensation offered by the company. In a laboratory study, Chatterjee et al. (2019) showed that varying reactions to a data breach depend on whether the customer affected felt anger versus fear following the notification. Chen and Jai (2021) found that customers in a hotel's loyalty program did not differ from their non-loyal counterparts in terms of perceived vulnerability following a hypothetical data breach but did in (dis)trust toward the hotel that was breached. However, these findings were based on surveys and laboratory studies, and customers may fail to recall or disclose their actual behavior accurately. When asked, people tend to convey a judiciously privacy-wary approach, yet multiple studies of willingness to share information have found that their revealed choices are to share more than they state they would (Athey et al. 2017, Acquisti et al. 2020). In contrast to the extant literature studying the effects of privacy shocks in daily decisions, our empirical analysis measures change in behavior, not merely stated intent to alter behavior, in their aftermath.

Firms whose data were breached can suffer financially through loss of revenue but also via punitive measures like monetary fines (Romanosky et al. 2014, Wang et al. 2019). Yet firms can mitigate or even reverse reputational damage through their reaction in the breach's immediate wake; some respondents so highly value prompt notification that there is sometimes an increase in valuation following a breach (Ablon et al. 2016, Chen and Jai 2021). Reactions to a data breach are hardly limited to a firm's current customers. Zhong and Schweidel (2020) analyzed online conversations and found that references to a breached company were often initiated by people who had never discussed it before; moreover, negative sentiment following the breach was short-lived. Evidence suggests that firms can even benefit from negative buzz; for example, Han et al. (2020) found that negative attitudes toward a product or company (not specific to data breaches) can sometimes generate increased awareness and downstream purchase intent. All told, evidence suggests that a data breach announcement could, in some cases, entail longer-term net gain in awareness; that is, a positive effect 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, online sentiment, and/or laboratory 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 suggested 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, among 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.

In summary, the literature on the effects of data breaches suggests that users will differ in their reaction, with some even apparently increasing interaction with the breached company. Moreover, such differences can vary based on personal circumstances and perceived actions taken by the firm. Because it is of universal interest to encourage and enhance customers' protection, our primary goal is to develop a methodology to measure not only the (causal) effects of data breach but also one that allows the analyst 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.

1.2. Literature Review: Measuring the Effects of an Exogenous Shock

Understanding the range of reactions to an exogenous shock, including data breaches, is a quantitative problem that merits methodological 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, Goldfarb et al. 2022).

In addition, observable behavioral changes among customers 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 with the behavior that users would 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 literature. In the context of data breaches specifically, Janakiraman et al. (2018) compared changes in sales over time in breached versus non-breached channels, relying on the consistency of customers' trajectories. Specifically, the authors use a diff-in-diff-in-diff methodology, overcoming the challenges presented by a unique setting where only one channel was affected, leveraging the relatively consistent prebreach behavior of customers in both channels to engage in counterfactual analysis. By contrast, the data breach presented in this paper affected all users, and users' behaviors prior to the data breach were not consistent enough to assume that they would persist had the breach not occurred.

Beyond the example of Janakiraman et al. (2018), other studies exploring the consequences of data breaches relied mainly on public data or surveys. Owing to the broad availability and reliability of stock market data, measures of change in valuation were also presented for publicly traded firms (Acquisti and Varian 2005, Martin et al. 2017, Amir et al. 2018). However, the validity of such measures relies on a number of assumptions, for example, consistent behavior of the corpus of customers among channels, lack of spillover effects between channels, or that the company be public in the first place. As mentioned earlier, other methods relied on surveys or laboratory studies, and therefore, 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) noted several empirical analyses, along with surveys and laboratory studies. Of the methods presented in this review paper, measures of aggregate change to sales, compared with other brands, were proposed, as well as event study methods. To reiterate, all these methods either assume that there are comparable products, assume consistent behavior and limited spillover effects, or provide only aggregate analyses.

In practice, data breach announcements are rarely confined to those customers whose data were compromised. For the particular breach analyzed in this paper—and many others, for example, Equifax, Uber, Target, the LGBTQ website Atraf, Ashley Madison, FriendFinder, AOL, and Yahoo, to name but a few—the announcement generated headlines. Even assuming that noncustomers can serve as controls is problematic; once the cat is out of the proverbial bag, finding an uninformed group with characteristics similar enough to those affected directly can be a practical impossibility. Similarly, assuming otherwise-consistent customer behavior, had the data breach not occurred, is unsustainable. In order to estimate the effect of the data breach accurately, predictable fluctuations and trajectories in users' actions must be accounted for when estimating counterfactual behavior.

Exogenous shocks are hardly limited to breaches but extend to public policy changes, corporate gaffes, and even natural disasters. Such events may require changes to firms' product offerings or distribution patterns, among other remedies. Methods such as Bayesian or Generalized Synthetic Control (Xu 2017, Kim et al. 2020) are suited to this domain because they allow the creation of a "similar-enough" control group when there are several close (but not perfectly comparable) control groups to begin with (e.g., other states or countries that have not yet regulated or have not been impacted). Regression discontinuity methods (Calonico et al. 2017, 2019, Cattaneo and Titiunik 2021) have also proved useful in estimating temporal changes sans control group. We use these methodologies as robustness checks, discuss their limitations, and assess their substantive similarity to the Temporal Causal Inference methodology developed here.

Similar methods for estimating changes in users' behavior also appear in the domain of "Buy Till you Die" (BTYD) models for customer base analysis. Specifically, Gopalakrishnan et al. (2017) proposed a cross-cohort vector changepoint model to better understand usage trajectories for newer users with little available transaction data. Although this method could potentially be extended to estimate the effects of an exogenous shock, among its main benefits is the ability to detect a *latent* regime change and as such typically requires sufficient post-change observations.[5] Other options for estimating time-varying effects with BTYD models include Bachmann et al. (2021), which requires the events to be either seasonal or to vary, and repeat, across customers.

Therefore, our overarching goal is 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. 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 Causal Forests in order to assess the heterogeneous treatment effects attributed to the announcement of the data breach.

2. Data Description

The focal data come from a matchmaking website aimed at those seeking extramarital affairs. 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, identifiable information for all paying members, and preferences for affair types, among other potentially socially embarrassing elements. The announcement was highly publicized by major media outlets in the United States and abroad, and the website itself made multiple announcements to their users and the general public. Therefore, it is reasonable to assume that news of the breach reached the entirety of the website's user base in short order.[6]

Our collaboration with the website afforded detailed user behavior records as well as de-identified profiles of all paying members. The data window for this analysis commences approximately six months before the announcement of the breach, through three 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.

2.1. Behavioral Data

We observe searches made and messages sent by users, as well as one specific "deliberative" action: deletion of photos.[7] Our sample consists of all paying male users from the United States who had joined the website one to six months before the breach was announced and had at least one "activity" (searches, messages, etc.; anything beyond the mere creation of a profile) on the website before the breach was announced. The relatively long span of join dates allows for an account of changes in activity patterns before the breach—for example, satiation, attrition, and/or other trends and fluctuations in individual-level behavior—and to also explore whether users who joined the website earlier differ in their reactions from those who joined later. Each user is assigned a unique ID that does not change over time, allowing us to view all activities users engaged in throughout the data window.

The focal website used a so-called "freemium" model, where one can join for free and enjoy limited functionality. Whereas women obtain essentially unlimited access for free, men must pay for credits in order to become "full" members; only then can they contact other members, until they run out of credits,[8] but any user on the website can browse anyone else's content, including other users' photos, unless they explicitly make their photos private. Because of this feature of the website and the nature of the breached data, our analysis and statistical estimates pertain to male users that had paid at least once for credits, because all such users had to provide their full name and billing address in order to process payment. This is not so for users who never paid to be active members, many of whom were pseudonymous, and so we limit our purview to legitimate, accurately recorded male users. Importantly, these users were informed, on the day of the breach announcement, 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 approximately 52,000 users, apportioned into 24 weekly cohorts (groups of users who join in the same week; mean cohort size = 2,174; SD = 296; min = 1,850), based on week of joining. Figure 1 illustrates the percentage 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 in each cohort initially increases and then gradually decreases. The breach occurred 27 weeks after the first cohort joined, and our data window extends to three full weeks after the breach was announced (and before the data were made public).

2.2. User Profiles

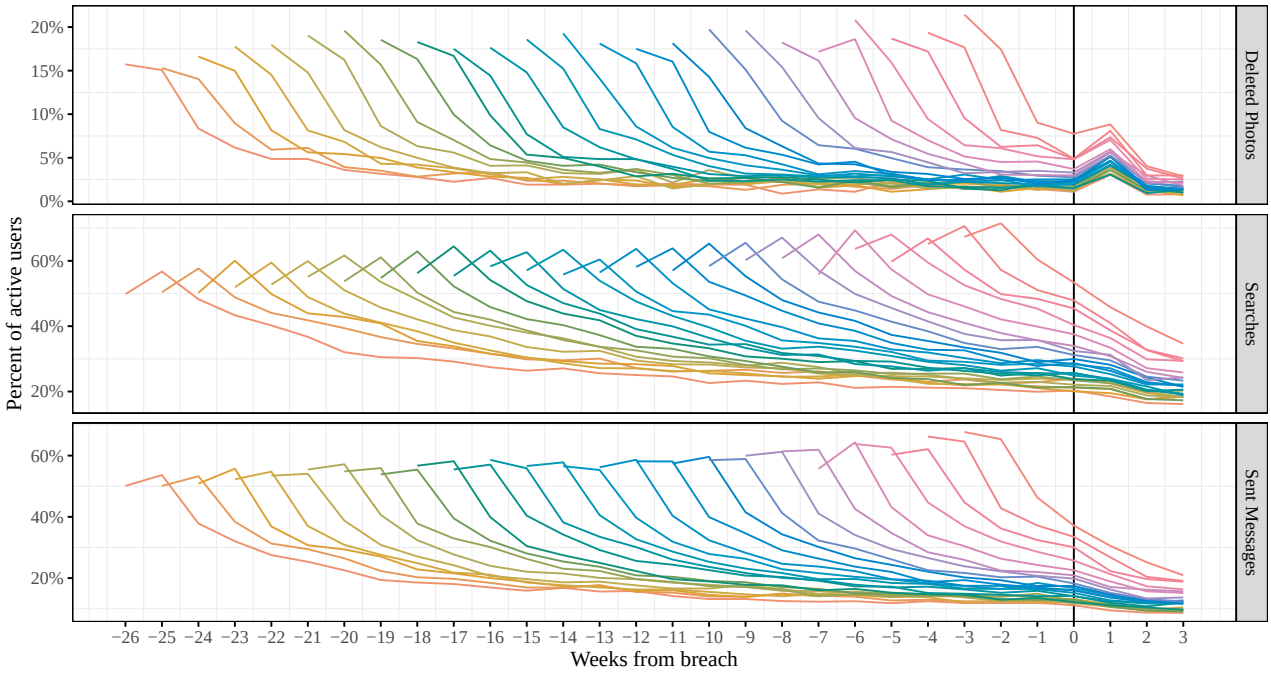
Upon joining the website, users provide a full profile, which includes gender, marital status, date of birth, height, weight, ethnicity, and several other sociodemographic covariates. Among the key specific covariates used in the analysis are marital status (attached or single, in a 67%−33% split)[9] and a binary indicator of whether the user's public profile includes a photo. Users can choose whether all other users can see their photos or only those who reciprocally shared their photos. Of all users, 85% had public photos, and 15%, referred to as "private" users, required reciprocity in sharing. The default was, as suspected, not being private. These covariates would be expected to correlate with an important latent element of public disclosure: that some users had "more to lose" than others.[10]

3. Methodology

In this section, we detail the main methodology used to identify treatment effects. It consists of two parts: Temporal Causal Inference, which involves the construction of treatment and control groups (even though all users were effectively "treated"); followed by Temporal Causal Forests, which uses the constructed control and treatment groups to nonparametrically estimate individual-level treatment effects, which can then be used in subsequent analyses. Later, we will show that these substantive results are robust both to other causal identification strategies and to alternative estimation methodologies. The combination of TCI and TCF, however, allows us to identify individual-level treatment effects fully nonparametrically and thereby assess heterogeneity in reaction to the data breach.

An outline of the overall methodology, to be described in detail shortly, is as follows:

Figure 1. (Color online) Percentage of Users Engaging in Each Activity (Deleting Photos, Searching, Messaging), For Each Cohort (Coded by Color), as a Function of the Number of Weeks Relative to the Breach Announcement (Vertical Lines)



1. Step 1: **Temporal Causal Inference:** For each cohort, acting as "treatment group," select cohorts that joined before this cohort's join date. They will be the applicable "control group," until they were exposed to the treatment: the data breach itself. Repeat this for all cohorts that have enough cohorts who joined before them. In essence, this compares control and treatment groups based on their trajectories on the website. Further details on this procedure appear in Section 3.1.

2. Step 2: **Temporal Causal Forests:** Match the trajectories of each treatment group with their respective control group, using Causal Forests. This nonparametric matching will be based on the activities each user made each week prior to the treatment. The treatment effect will be estimated individually, based on the difference between the predicted likelihood to be active sans treatment and the observed outcome with the treatment. We then estimate the average treatment effect and explore sources of heterogeneity.

We next detail these two key steps in the causal estimation methodology.

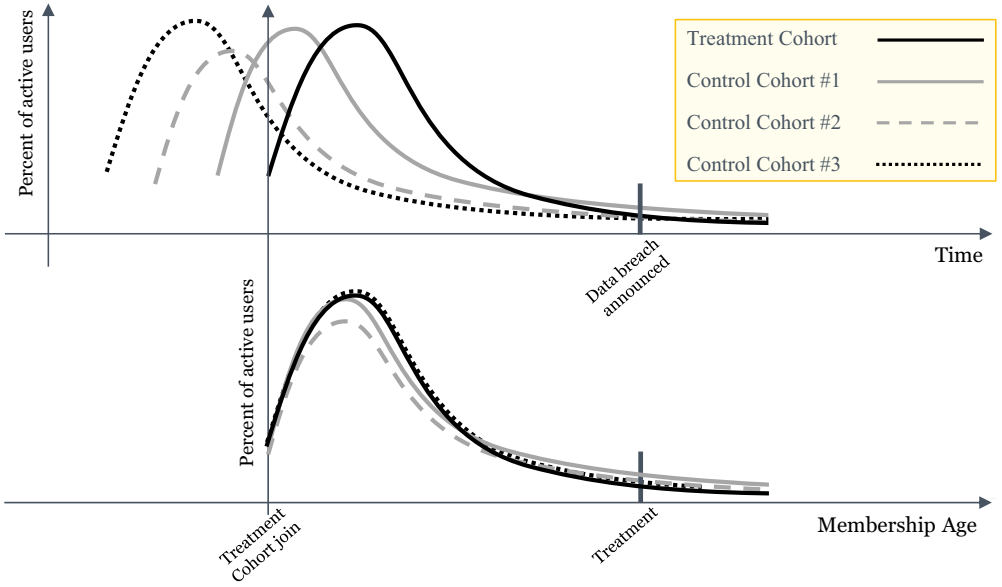
3.1. Temporal Causal Inference

Our goal is to estimate the effect of the data breach on the probability that a user will be active on the website. We refer to the announcement of the breach as a single, exogenous *treatment* whose effect we wish to estimate. As detailed earlier, the main challenge is that there is no clear control group, because—due to extensive publicity and the nature of the public announcement—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, sometimes referred to as "tenure" on the site to avoid confusion with actual user age). Site activity varies substantially as users spend more time on the website; 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 still others have distinct patterns of activities throughout their membership lifetime. That is, although the breach "hit" all users, it did so at different points in their experience and consequent activity pattern on the site.

Although users differ in their trajectories, having a relatively large number of users in each of the 24 cohorts permits excellent matching of users' trajectories 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 more 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 earlier in their tenure on the site compared with the control. In particular, let J_T denote the cohort of users who joined T weeks before the data

Figure 2. (Color online) Illustration of Temporal Causal Inference, with 1 Treatment and 3 Control Cohorts



breach was announced. For this cohort, we observe $T+3$ weeks, where the last 3 weeks are after the breach announcement. Let groups $J_1, \ldots, J_{T-1}$ be all the cohorts that joined at least one week before J_T. From these cohorts, we use their activity since joining the site and until the data breach occurred, or until $T+3$ weeks since joining (whichever comes first),[11] and form a control group J_T^C.

It is critical to note that all users in the control group were also affected by the breach. However, they were affected later in their tenure on the website (i.e., time since joining). For this control group, J_T^C, the time of (not receiving the) treatment will still be T, and the weeks to follow (which were all before the breach) will enable predicting what would have been the (expected) behavior of the treatment group had the breach not been announced.

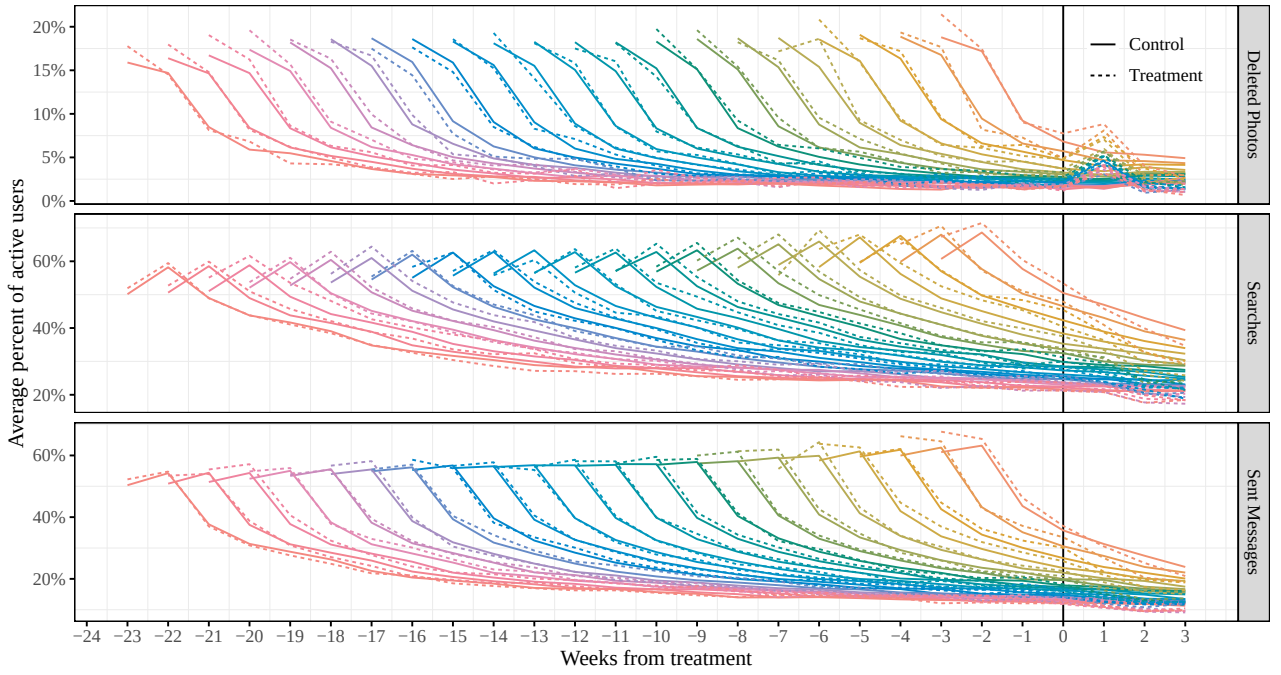
We illustrate the construction of these groups in Figure 2. The upper panel illustrates the average number of activities for one treatment cohort and three earlier cohorts over time; the earlier cohorts are formed into a single control group. The lower panel illustrates that the groups are comparable, except that the control group was not yet exposed to the breach (note that in the lower panel, the x-axis variable is number of weeks since joining the website, i.e., not calendar time). For the control group, we use only the data *up to the treatment*, or up to the available timeline for the treated cohort, as described above.

For each of the cohorts who joined in a specific week one to six months prior to the data breach, we employ all of the users who joined at most eight 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 times since they joined the website) and that almost all cohorts serve as a treated group. The exception is for the first three and last three cohorts that act either only as control or only as treated groups, but not both, because of their location at the edges of our data window.

Figure 3 provides a visual comparison between control and treatment groups, as constructed by Temporal Causal Inference (TCI),[12] whereas Figure 4 relays the average number of activities for each of the three weeks prior and post-announcement, across all cohorts, because they are divided into control and treatment groups. This figure depicts a sort of "model-free evidence" for the change in the probability of being active in each of the activities on the website in the weeks following the announcement. This model-free evidence suggests a notable increase in the likelihood of deleting photos immediately after the data breach was announced and a decrease in searches and messages. We note that searches seem to have decreased only in weeks 2 and 3 after the data breach, whereas Figure 4 suggests that the likelihood to message was reduced already in the first week post-breach. It also underscores a potential limitation in the ability to estimate the change in the likelihood of deleting photos in weeks 2 and 3 following the data breach because the control group had not followed a monotonic trajectory. We revisit this issue and ways to overcome it in Section 4.1.

3.1.1. Causal Identification Assumptions with Temporal Causal Inference. As stated, one major obstacle in identifying the consequences of an exogenous shock is fashioning a proper control group. We follow the so-called

Figure 3. (Color online) Percent of Active Users per Control/Treatment Group, as a Function of Week from Treatment

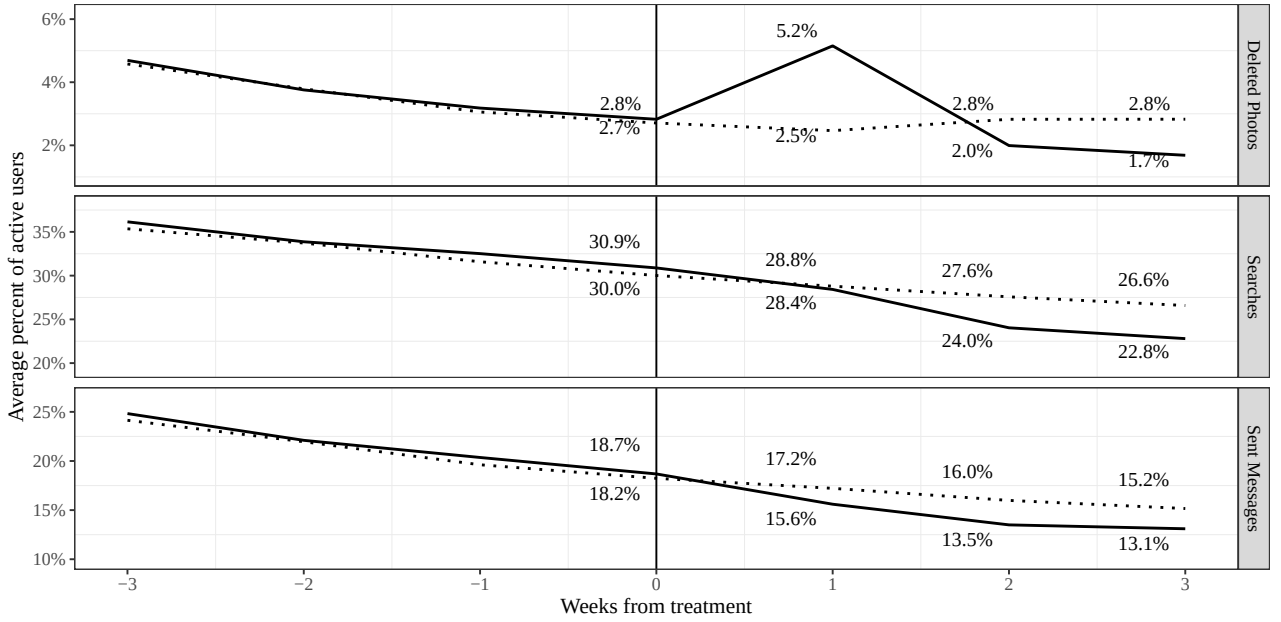


Note. Solid vertical lines indicate breach announcement; colors indicate week of joining for each of the 21 treatment groups.

"potential outcomes framework" (see, e.g., Rubin 2005) and in the Electronic Companion show that all the necessary assumptions for causality claims are met.

Specifically, EC.1.1 addresses concerns of interference between treated and control users, first because of the real-world time gap between them, and secondly because of the stability of the site (which at the time of announcement was in a period of sustained growth). Moreover, given that the site exists to introduce dating partners, network effects may affect two treated or two control users before and after the treatment. However, we show that this does not harm the ability to measure treatment effects; if users of the treated groups changed their behaviors in ways that affected others in their

Figure 4. Average Percentage of Active Users Across All Cohorts; Control Group (Dashed), Treatment Group (Solid)



group, then this is an important part of the treatment effect and should be taken into consideration in its assessment.

We further describe in EC.1.2 that, because of the construction of the control group as comprising several cohorts, we can alleviate or at least substantially reduce concerns about seasonal artifacts. Using both Granger Causality (Granger 1980) and Kolmogorov-Smirnov (Massey 1951) tests (EC.1.2), and via a series of Placebo tests (Appendix B.2), we show that the cohorts are similar in their pretreatment behavior and empirically test that two fundamental assumptions are met: parallel trends and conditional independence. Specifically, the Granger Causality and Kolmogorov-Smirnov tests compare trends (prior to the treatment) of the control and treatment groups, finding no significant differences. A series of placebo tests focused on five weeks before the actual treatment time yield null or negligible results, that is, correctly failing to infer treatment effects prior to the data breach.

Lastly, in EC.1.3, we discuss the testing of the overlap assumption; all users were treated eventually, but at different points in their tenure on the website. That is, although all users evidently had some probability of being treated, we further test their likelihood to have been treated empirically using Causal Forests. We also address the exogeneity of covariates; being in the (role of) control or treatment group is unrelated to behavior prior to the treatment (or lack thereof). It must be emphasized that no one prophesied this particular data breach, and there is no basis for believing that concern about a data breach should differ across treated and control users. That said, if there are indeed any such differences, most users in TCI act both as control and as treatment users, and we shall in addition control for other differences using the next step in our analysis, Causal Forests.

To summarize, Temporal Causal Inference aids in fashioning appropriate control groups to measure the treatment effect on the exogenous shock of the data breach. In the next section, we introduce Causal Forests (CF), a nonparametric method that will further allow us to find a matching control for each user in our treated groups. Temporal Causal Forests—the combination of TCI and CF—will allow measurement of treatment effect heterogeneity, which can in turn be related to individual-level characteristics.

3.2. Temporal Causal Forests

We use the control and treatment groups created via TCI and estimate individual treatment effects using a nonparametric, forest-based method: Temporal Causal Forests. The magnitude, direction, and significance of these effects will be further assessed by a wide variety of parametric and semiparametric models as well as dedicated simulation studies to verify parameter recovery when the data-generating process is known. These additional analyses will be described in Sections 5–8.

3.2.1. Construction of Temporal Causal Forests.

We observe three to 24 weeks of activities prior to the treatment for each cohort, the variation stemming from time of their joining the site. This timeline is compared with that of similar users in the control group to estimate the probability of being active following the treatment had the breach not occurred. In order to do so, we use a forest-based method, specifically, Causal Forests. It is important to note that Temporal Causal Inference and Causal Forests comprise two separate, sequential procedures; TCI generates the control and treatment groups and, together with Causal Forests, generates heterogeneous treatment effects by estimating a user's probability of being active if the user had been in the opposite group. Simply put, the difference between the estimated likelihood of being active and the observed activity provides an estimate of the effect of the breach. The second step, using Causal Forests, 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 adopt the Causal Forest (CF) (Wager and Athey 2018) model using a Generalized Random Forests (GRF) implementation (Athey et al. 2019). A recent, extensive simulation study found Causal Forests in this implementation to perform exceptionally well under all tested settings (Knaus et al. 2021). The robustness checks in Section 5 will further showcase how the construction of control and treatment groups with Temporal Causal Inference can be used in conjunction with other causal inference methodologies. Although robust, we find that the pairing of TCF was especially accurate in estimating both average and individual treatment effects, even when the latter deviated considerably from any "named" distribution toward which the analyst might regularize a priori. As such, our main analysis refer nearly exclusively to TCF.

3.2.1.1. Causal Forests (Implemented Using GRF).

Causal Forests and the specific weighing mechanism behind GRF are detailed in EC.2. Here, we describe the construction of the trees and forests specific to our application.

In each tree, we use as the set of features (covariates) $x = X_i$, the timeline of each user in the control and treatment groups. The timeline is a vector of binary variables: whether the user engaged in each one of the focal activities (described in detail in Section 4) in each week since joining (NULL if not joined yet).

Let $W_i \in \{0,1\}$ be the treatment assignment of user i and $Y_i = Y_i(W_i)$ be the observed outcome (being active or not in any particular activity in either of the three

weeks after the data breach). 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

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

In each tree, we recursively partition the users into two leaves. Usually, the splitting rule is set to minimize the sum-of-squared error between the observed and predicted outcome (further explained in EC.2). However, in causal inference, we do not observe both $Y_i(1)$ and $Y_i(0)$ but rather only one of them. For Causal Forests in the GRF implementation, the splitting rule is set to maximize the heterogeneity in the treatment effects between the leaves, which is equivalent to minimizing the sum-of-squared error between the observed and predicted outcome sans treatment (see Proposition 1 in Athey et al. 2019).[13]

The maximization of heterogeneity is based on both pseudo-outcomes and the propensity to be treated. Let the pseudo-outcome be

$$\rho_i = \frac{((W_i - \overline{W}_P)(Y_i - \overline{Y}_P - \hat{\beta}_P(W_i - \overline{W}_P)))}{\mathrm{Var}_P(W_i)},$$

where $\overline{W}_P$, $\overline{Y}_P$ are the averages of the propensity to be treated and the predicted outcome, respectively, taken over the parent node P, and $\hat{\beta}_P$ is the least-squares regression solution of Y_i on W_i in P. $\mathrm{Var}_P(W_i)$ is the variance of the treatment in parent node, P:

$$\mathrm{Var}_P(W_i) = \frac{1}{|\{i : \boldsymbol{X}_i \in P\}|} \sum_{\{i:\boldsymbol{X}_i \in P\}} ((W_i - \overline{W}_P) \otimes (W_i - \overline{W}_P)^T)$$

The splitting rule is then calculated along the gradient of the mean difference with the pseudo-outcomes. Specifically, the parent node will be split into two leaves $\{C_1, C_2\}$ that maximize

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

This splitting criterion is applied recursively, thereby generating a tree and, later, a forest. The weighing mechanism specified in the description of GRF (as per EC.2) then takes place.

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

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

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

Then, center the outcome and treatment $\tilde{Y}_i = Y_i - \hat{y}^{(-i)}(\boldsymbol{X}_i)$ and $\tilde{W}_i = W_i - \hat{w}^{(-i)}(\boldsymbol{X}_i)$, where the $(-i)$ superscript denotes out-of-bag estimates of $\hat{y}$ and $\hat{w}$, computed without the ith observation. Using these quantities, forests can be run on the pair $\tilde{Y}_i, \tilde{W}_i$. Finally, after constructing the forest and retrieving the weights $\alpha_i(\boldsymbol{x})$ for an out-of-bag observation with covariates $\boldsymbol{x}$, we can estimate the treatment effect via (Athey and Wager (2019), Equation (4)):

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

As stated earlier, Wager and Athey (2018) derived asymptotics of $\hat{\tau}(\boldsymbol{x})$ and thereby estimated individual treatment effects' mean and variance, $\tilde{\sigma}_\tau(\boldsymbol{x})$. The variance is estimated using the infinitesimal Jackknife method, also known as the nonparametric 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 details on these methods.[14]

In each Temporal Causal Forest:

1. Estimate nuisance parameters $\hat{y}(\boldsymbol{x})$ (the probability of being active) and $\hat{w}(\boldsymbol{x})$ (the probability of being treated) for all users. In our main analyses, where all cohorts are estimated jointly, this is done using the built-in regression forests within Causal Forests. In robustness checks, when estimating each cohort separately, this was done using Local Linear Forests, further described in EC.2. Parameters were tuned via default settings, which bootstrap over the available parameter space and optimally find suitable scaling parameters.

2. Build a Causal Tree on GRF, which classifies users from the control and treatment groups, based on their set of features, $\boldsymbol{x} = \boldsymbol{X}_i$. 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 each one of the activities each week[15]) before the treatment ($\boldsymbol{X}_{it}$). We construct 4,000 honest trees.[16] The parameters used in creation of the trees are estimated using the default tuning so that the optimization of the parameters is carried out by bootstrapping.[17]

3. The results from running the TCF include individual estimates of treatment effects; each user in a treated group has nine treatment effects, $\hat{\tau}_i$, estimated: for each of the three types of activities (sent messages, searches, and deleted photos) and for each of the three weeks after the announcement. Each of these nine treatment effects is a measure of the change in probability of being active in this particular activity for that week.[18]

4. In addition to the individual mean treatment estimates, Causal Forests' ability to recover asymptotics also enables us to estimate the variances of each of these nine

means; these are indirectly based on the ability to match control and treated users and will be denoted $\sigma^2_{\tau_i}$.

We create a single Temporal Causal Forest for all treatment groups, along with their respective control groups, for each week after the treatment $p \in \{1,2,3\}$ and for each type of activity (sent messages, searches, deleted photos).[18]

3.2.2. Population Mean and Variance. To estimate both the mean treatment effect across the population and the associated variance, we proceed as follows; let $(\hat{\tau}_i, \sigma^2_{\tau_i})$ 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:

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

The variance around this estimation is constructed from both the uncertainty (variance) around each individual estimate $(\sigma^2_{\tau_i})$ and the uncertainty around the mean of all individual estimates:

$$\mathrm{Var}(\overline{\tau}) = \frac{\sum_{i \in treated} [\sigma^2_{\tau_i} + (\hat{\tau}_i - \overline{\tau}_{treated})^2]}{N^2_{treated}}$$

4. Results

Here, we consider the main objects of inference arising from the causal forests: the individual-level treatment effects and their distributions. To do so, we first examine (mean) treatment effects—across outcomes, over time, across individuals—subsequently accounting for individual-level covariates and various forms of robustness.

4.1. Average Treatment Effect

As per Table 1, the announcement on deleted photos suggests a substantial (average) increase; 2.8% more of the users (i.e., double the baseline) deleted at least one photo immediately after the announcement (week 1). This appears reasonable, consistent with users trying to "cover their tracks" after realizing that their personal data were no longer secure on the site. Although hardly a remedy for their data security loss, it was at least a protective action under their own control (and one we can observe) to reduce the chances of further identification. While the data were in the hands of the hackers, images on the website were visible to almost every user who joined, even without being a paying member, and therefore, deletion of photos was an immediate first step that could have offered protection, at least to some degree. In the following weeks, there is a substantial decrease in the probability of deleting photos compared with before the announcement; this suggests that those users wishing to delete their photos did so immediately after the announcement, leaving fewer photos to delete in subsequent weeks.

Table 1. Percent Change in Likelihood of Being Active: Population Mean and Standard Deviation

		Week 1	Week 2	Week 3
Deleted photos	Mean	2.805	−0.166	−0.300
	SD	(0.012)	(0.001)	(0.001)
Searches	Mean	−0.289	−3.363	−3.369
	SD	(0.005)	(0.012)	(0.011)
Sent messages	Mean	−1.426	−2.425	−1.921
	SD	(0.006)	(0.012)	(0.005)

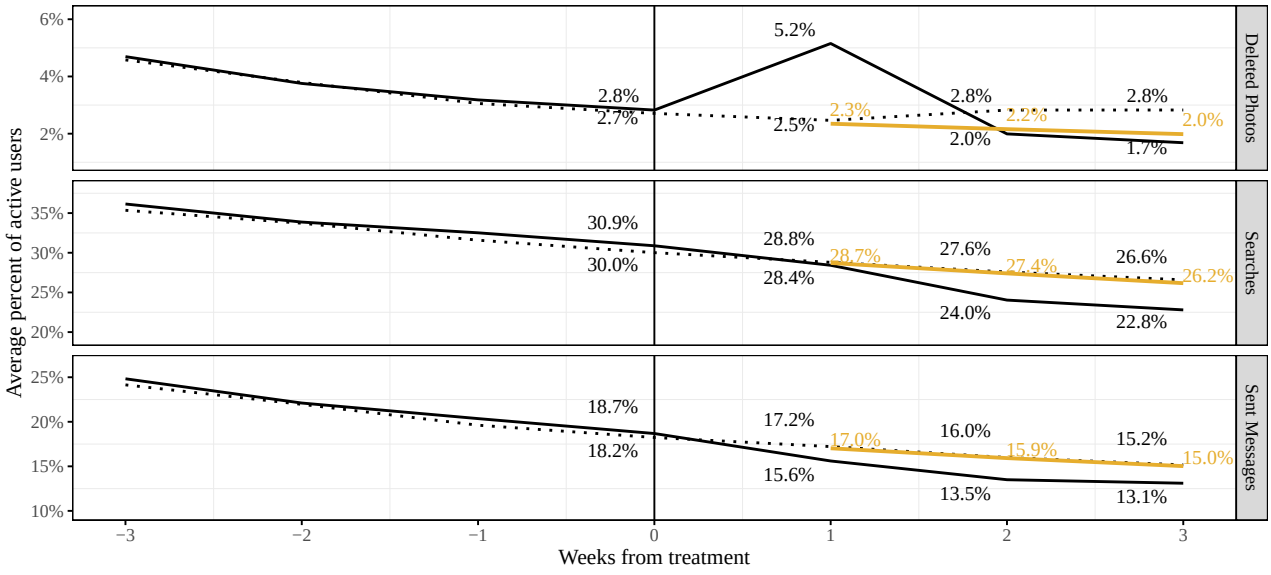
For searches and sent messages, users (on average)—in contrast to photo deletion—decreased both activities for all three weeks post-announcement. For searches, the effect in activity probability manifests mainly in the second and third week after the announcement (around 3.4% decrease in each week). For sent messages, results suggest some attenuation (i.e., "less decrease") in week 3 compared with the week before (2.4% in week 2 vs. 1.9% in week 3). Notably, this initial attenuation of the treatment effect is measured while controlling for possible trends that would have emerged sans treatment, because we nonparametrically match users from the control and treatment groups based on their trajectories on the website prior to the treatment (or lack of it for control users). Such plateauing/waning over time echoes Rosati et al. (2017), who found a short-term reduction in bid-market price following a data breach but no long-duration effect. That said, our three-week post-breach data window does not allow verification beyond that time period.

Model-free evidence for TCI appears, as previously, in Figure 5 but is supplemented by estimated counterfactuals by causal forests. These contours, taken together, suggest that TCFs accurately track the percent change in probability of being active; the estimated counterfactual closely follows the observed control group's behavior. Although there appears to be one anomalous deviation, deleted photos in weeks 2 and 3, this occurred because the cohorts used as control group increased their activities in a way at odds with past behavior, which was monotonically decreasing. Because the counterfactual (via TCF) is calculated based on individual behavioral trajectories, it nonparametrically extrapolates the downward trend, as opposed to the idiosyncratic increase seen in the control group. The treatment effects of Table 1 are then estimated relative to this counterfactual.

4.2. Treatment Effect Heterogeneity (Across Users)

As previously noted, users may vary in their reactions to the data breach. The distributions of the individual treatment effects estimates are displayed in Figure 6.

Figure 5. (Color online) Percent of Active Users per Control (Dashed Line) and Treatment (Solid) Groups, Along with the Estimated Counterfactual from Temporal Causal Forests (Orange)

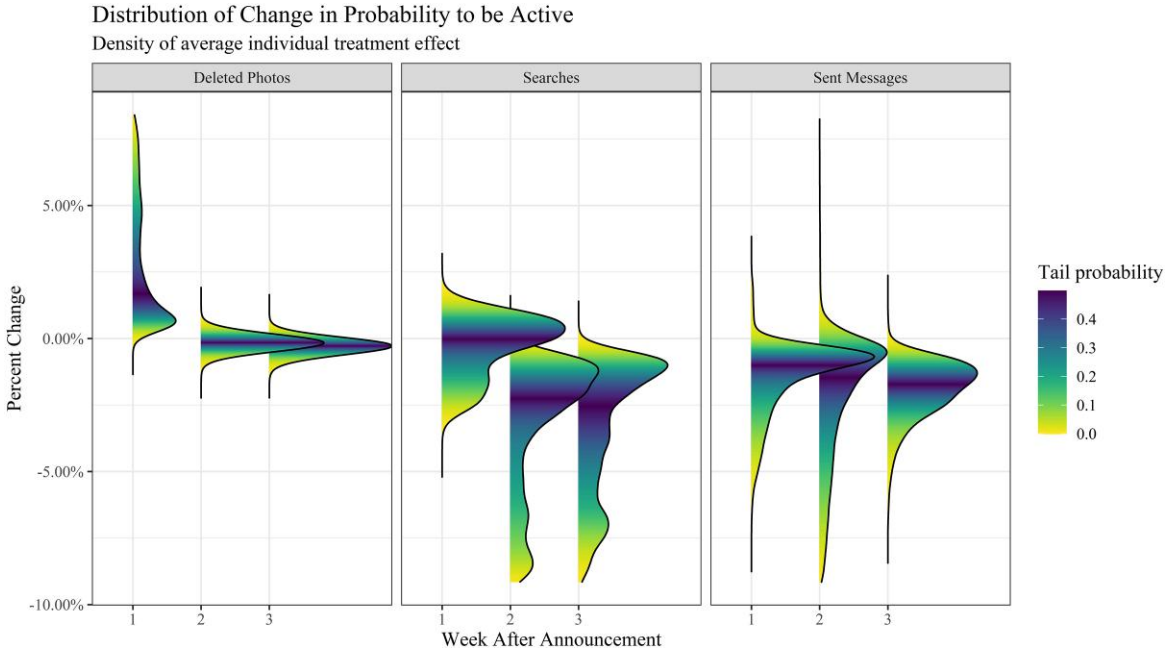


The forest-based approach affords for nonparametric regularization, and this flexibility is apparent in the non-Gaussian density plots. This suggests that other methods that impose specific parametric regularizations, with or without covariates, may entail misspecifications or be sensitive to outliers (as discussed in Section 5). We proceed by first examining these density plots and then relating them to covariates using a variety of methods.

4.3. Sources of Heterogeneity in the Effect of the Breach

As documented in the literature review, users may have varying reactions to data breaches, owing in part

Figure 6. (Color online) Distributions of the Heterogeneity in Treatment Effects; Mean Individual Changes in the Probability of Being Active in Each Activity on the Website, as Estimated by Temporal Causal Forests



Note. Each panel is for a different activity, each column is for different week following the announcement of the data breach, and colors indicate tail probability.

to different expectations of websites' duty to protect their personal information. Moreover, the breached data could reveal illegal or socially proscribed activity, such as an active search for an extramarital affair.

A specific example in our data setting concerns whether married people would be differentially affected by the breach because they have, *ceteris paribus*, more to lose compared with singles. In such cases, heterogeneity may be partly explicable by users' willingness to be more "public" in their profile, assessed by whether the user had a viewable (public) photo on the website.[19] Users with public photos run the risk of revealing their identity even before the breach (e.g., depending on how identifiable they are in their photos), potentially because they entrusted the website's security protocols.

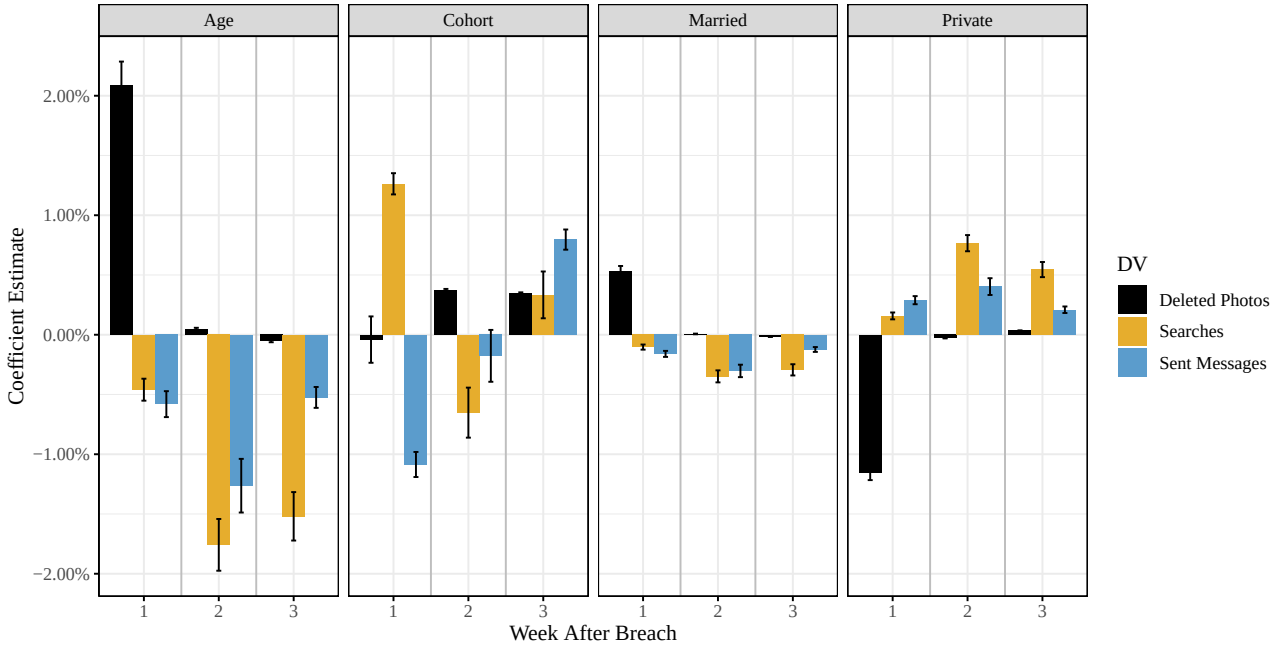
Although Figure 6 visually depicts the substantial heterogeneity in treatment effects, it does not relate them to user actions or characteristics. To assess observed sources of heterogeneity in the response to the breach announcement, we complement Temporal Causal Forest results with a variety of models, each "regressing" either individual treatment effects (or, separately, observed outcomes) on other individual-level covariates. For the purposes of illustration, the simplest such case would be a linear (in parameters) specification with no interactions, for example, for user i with treatment effect $\hat{\tau}_i$ (e.g., for searches in post-breach week 2):

$$\hat{\tau}_i \sim \beta_0 + \beta_A Age_i + \beta_C Cohort_i + \beta_M Married_i + \beta_P Private_i \quad (1)$$

For ease and consistency of interpretation, all key independent variables—*Age*, *Cohort* (higher values mean the user joined in a later week), and binary indicators of whether the user was *Married* and was *Private* (had no public photos on the site)—are mean-centered so that the intercept corresponds to the "centroid case" within the data and will be relatively stable across covariate specifications. We start with these simple linear models and then explore a variety of more complex ones. Note that this is fully analogous to "Level II" of a hierarchical model, except that the dependent variables here are the individual treatment effects, as provided by Temporal Causal Forests.

The "β" coefficients[20] presented in Figure 7 capture marginal effects of *deviations* from the average effect of the data breach for those who are, respectively, older (higher *Age*), "newer" on the website (higher *Cohort*), *Married*, and *Private* (had no public photo). The coefficients indicate that newer users on the website (higher "Cohort") had more extreme reactions in terms of messages but were quicker to bounce back toward typical trajectories (positive coefficient for messages in week 3). For searches, newer users had slower reactions, but also here they were coming back quickly, already in the third week after the announcement. For searches and messages, results also suggest that married people reduced more of their messages and searches compared with 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 a smaller decrease (smaller treatment effect, as reflected in a positive coefficient) in terms of the probability of being active in searches and messages than those who chose to have public photos on the website.

Figure 7. (Color online) Coefficients and 95% CIs for Sources of Treatment Effect Heterogeneity for Each Type of Activity [Age and Cohort Based on 75% vs. 25% Interquartile Ranges, as per Main Text].



In terms of photo deletion, married users deleted more photos immediately after the announcement, whereas private users deleted fewer photos than public users. This might be due to the fact that private users did not understand that allowing only reciprocal sharing of photos does not mean they are protected from other users who might join or from the data breach itself.[21] This could well stem from a false sense of privacy (Brandimarte et al. 2013), one where, ironically, those who cared most about their privacy (e.g., changed their settings to disallow public photos available to all[22]) were also the least, and last, to protect their privacy in what in the end was only way they could have.

5. Alternative Models and Robustness Checks

Although Causal Forests allow for fully nonparametric regularization, it is important to bear in mind that no one causal inference method provides a "ground truth." As such, we estimate a wide variety of alternate models to assess convergent validity regarding the magnitude and direction of the treatment effects and how they relate to covariates. We will, generally speaking, present four sorts of analyses based on the dependent variable itself (causal forest results vs. original data) and its scale properties (continuous vs. binary/binarized, individual level vs. cohort-level analysis). We also test the basic identification through Temporal Causal Inference by including methods that do not rely on it.

A list of the types of models estimated is as follows, broken out by link function, nature of the DV, and whether TCI is formally carried out (i.e., via treatment and control groups), each explained more fully in the sequel and in Table 2. (Full estimation results for all models are posted for review at this https://osf.io/ekjgz/?view_only=1d9623a1262c48edb54870fc1b51666c.)

The first model, "OVERALL" (M01), contains no covariates, so it merely reproduces the mean forest effects of Table 1; we will use this throughout as a set of "base" results against which others can be compared.

The next group, {LM, RLM, GAM} (M02-04), are, respectively, Linear, Robust Linear (Li 1985), and Generalized Additive (Hastie and Tibshirani 1990) Models, with first-order interaction terms (higher order were *ns*, omnibus). These, taken as a group, explore robustness to relating DVs to covariates (Age, Marital status, Privacy preferences, and Cohort), to avoid parametric/linearity assumptions for covariate effects.

The third group, {MultiForest, MultiMulti} (M05, M06), are alternative methods for running TCFs estimating multiple DVs and multiple weeks at a time.

The fourth group (M07-11) contains a variety of classic and recent causal inference approaches with varying assumptions: stand-alone methods like Diff-in-Diff (DiD), Regression Discontinuity in Time (RDIT), Generalized Synthetic Control (GSynth), and Bayesian Synthetic Control Method (BSCM). Note that both types of RDIT models show robustness to the construction within the Temporal Causal Inference method, because they do not require the creation of control and treatment groups.

The last group (M12-15) avoids the forest estimates' presumed scaling properties, either binarizing TCF results or using the original binarized data directly; owing to their logistic link, their coefficients will be on a different scale from models in the other groups. Full model descriptions, specifications, and results appear below and in the Electronic Companion.

6. Analyses and Robustness of Causal Forests Results

In this section, we present two classes of results: various regression-based approaches to understand heterogeneity in individual-level treatment effects and two additional causal forest models that account for multiple dependent variables in a unified manner.

Table 2. Specifications of Main and Alternative Models

Model	Model	Link	DVType	TCI	Interactions
M01	OVERALL	Linear	Forest	Yes	- - -
M02	LM	Linear	Forest	Yes	Yes
M03	RLM	Linear	Forest	Yes	Yes
M04	GAM	Linear	Forest	Yes	Yes
M05	MultiForest	Linear	Forest	Yes	- - -
M06	MultiMulti	Linear	Forest	Yes	- - -
M07	DiDLinear	Linear	Individual	Yes	- - -
M08	RDITLinear	Linear	Cohort-level	No	Yes
M09	RDITNonParam	Linear	Cohort-level	No	- - -
M10	GSynth	Linear	Cohort-level	Yes	- - -
M11	BSCM	Linear	Cohort-level	Yes	- - -
M12	LogitForest	Logistic	Forest (binarized)	Yes	Yes
M13	RDITLogit	Logistic	Individual	No	Yes
M14	LogitLinear	Logistic	Individual	No	Yes
M15	LogitGAM	Logistic	Individual	No	Yes

6.1. Causal Forest Regressions: LM, RLM, and GAM

We first explore whether overall mean treatment ("intercept") and covariate effects are stable across regression approaches, where the DV remains individuals' treatment effects, and the model covariates (*Age*, *Cohort*, *Married*, *Private*) and their various interactions are related linearly (LM), robustly (RLM), and nonlinearly (GAM, for the first two regressors).[23] As apparent in Table EC.4, treatment effects estimates are highly stable across regression approaches, with essentially identical results for OVERALL, LM, and GAM and mild to moderate attenuation for robust regression (RLM). In other words, estimated values for the mean causal effect for all three DVs (Photos, Searches, Message) and weeks (1, 2, and 3) are robust to the inclusion of covariates and interactions (LM vs. OVERALL), nonlinearities (GAM vs. LM), and outliers (RLM vs. LM) and hew closely to those in Table 1.

6.2. Temporal Causal Forests with Multiple Outcomes

In estimating the individual treatment effects (and their variances) using Temporal Causal Forests, we constructed a separate forest per activity (photos, searches, messages) for each week (1, 2, and 3) following the treatment (or each of the treated cohorts), essentially a 3×3 design. A recent development in Causal Forests allows for estimation of the treatment effects of multiple outcomes jointly in a single forest: "Multiforest" for all three weeks (for each DV separately) and "MultiMulti" for all nine activities and weeks together. Results, which appear in Table EC.5, are consistent across methods, albeit with some modest difference in magnitude: MultiForest somewhat larger (e.g., photos in weeks 2 and 3) and MultiMulti very mildly attenuated.[24] In short, the "overall" treatment effects for all 3×3 cases are broadly robust to alternative forest specifications.

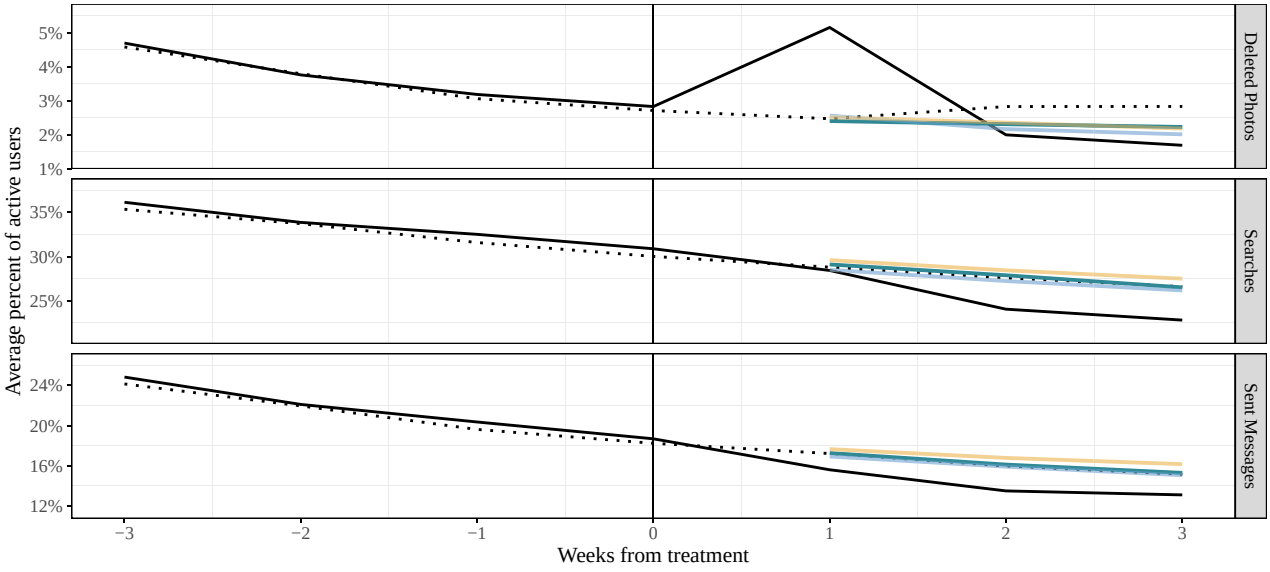
7. Alternative Causal Models: DiD, RDIT, GSynth, BSCM

It is critical not to simply "believe" causal estimates stemming from TCFs simply because they offer flexible nonparametric regularization. Here, we examine a variety of alternative approaches, including Diff-in-Diff, Regression Discontinuity, Generalized Synthetic Control, and Bayesian Synthetic Control, examining substantive similarity based on signs, magnitude, and significance of coefficients corresponding to treatment effects. All models are described briefly below. Further details, as well as comparative results for the four model types in question, can all be found in Electronic Companion EC.6, and visually below in Figure 8, and with additional models as per our forthcoming robustness analyses in Electronic Companion EC.8.

7.1. Diff-in-Diff (DiD) Model

We checked our main results against a "differences in differences" model ("DiDLinear"), following Janakiraman et al. (2018) in the domain of breaches, and numerous

Figure 8. (Color online) Average Percent of Active Users: Control (Dashed) and Treatment (Solid) Groups (Average Across Cohorts in Each Group), 3 Weeks Prior and Posttreatment, Alongside the Estimated Counterfactual Number of Activities for the Treated Group, Estimated Either with Bayesian Synthetic Control Method (Green, or Dark Gray in Grayscale), Generalized Synthetic Control Group (Yellow, or Light in Grayscale), or Our Proposed Chosen Method, Causal Forests (Blue, or Medium in Grayscale)



Note. Additional robustness analyses appear in Electronic Companion EC.6.

other 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 engaged in the focal activity at time t, μ_i is the individual fixed effect for user i, and $P_{it} = 1$ if the period t is posttreatment (or lack of treatment, for the control group) for user i, 0 otherwise; $Tr_i = 1$ if user i is in the treatment group, 0 otherwise, and A_{it} is the tenure (i.e., time since joining the site) of user i at time t; A_{it}^2 is included to allow for potentially diminishing marginal effects of tenure, as per the model-free illustration, and ϵ_{it} denotes error terms, with the usual zero-mean Gaussian (across users) assumption.[25]

7.2. Regression Discontinuity in Time (RDIT)

Another way to estimate treatment effects in time series data involves regression discontinuity methods, which have both parametric and nonparametric variants. Regression discontinuity in time (RDIT), in our context, utilizes the time trends of all cohorts to estimate the effect of the data breach on users' likelihood to be active in each type of activity. The "discontinuity" stems from the announcement of the data breach, and the model will account for its effects on the three focal activities after the breach was announced. Because of the nature of the model (i.e., including a single discontinuity), we estimate one treatment effect for each type of activity. RDIT is similar to a before-and-after study, but it allows for flexibly modeling the parameters governing the time trend prior to the treatment in order to assess the counterfactual trend and thereby estimate the average treatment effect. Further details on model assumptions and specifications, as well as accompanying results, appear in Electronic Companion EC.6.

7.3. Generalized (GSynth) and Bayesian (BSCM) Synthetic Control

As pointed out by Xu (2017), DiD models enact a strong "parallel trends" assumption; in the absence of treatment, the mean outcomes of the control and treatment groups should follow parallel paths. Although TCF overcomes this by nonparametrically matching via individual-level (parallel) trends, and because in the present case our empirical tests suggested that the assumption holds (see EC.2), it is nonetheless important to verify that focal substantive findings are robust to alternative casual inference approaches. To this end, we first enacted Generalized Synthetic Control (Xu 2017) (hereafter, GSynth, as implemented using `GSynth` R package). This affords several advantages over DiD, among them that GSynth can incorporate multiple time periods following the treatment and relax parallel trends by constructing a "synthetic control group" (i.e., a linear combination of trends of multiple users). As before, full details appear in Electronic Companion EC.6.

In addition to GSynth, we avail of the "Bayesian Synthetic Control Method" (BSCM) (Kim et al. 2020). This offers several benefits, including lack of a priori parameter space constraints, fully Bayesian inference, and intrinsic sparsity control. We estimated the BSCM model in Stan (Stan Development Team 2022), using both horseshoe and spike-and-slab priors, which gave nearly identical results; because the horseshoe prior ran three times faster, we present its results. Note that, here, analyses were limited to the cohort level because of the nature of the models; specifically, we had to limit the cohorts to those who joined at least four weeks prior to the breach. BSCM methodology allows the construction of a synthetic control group with varying degrees of similarity to control units, based on the selected prior, and has shown excellent recovery in simulations.

7.4. Robustness Comparison Checks for Alternative Causal Models

Comparative results for all four alternative causal models—DiD, RDIT, GSynth, BSCM—appear in Electronic Companion EC.6.1. In summary, although there is no "ground truth" among the causal inference methods, the proposed TCF-based method consistently lands near the center of the others and is never an outlier among the 3 (DVs) × 3 (Weeks) "design". It also has the benefit of providing individual-level treatment effects estimates that can be related to individual-level covariates, as in Section 4.3. Having surveyed mean treatment effects, in the next section we examine robustness for treatment effect heterogeneity, as well as to individual-level errors.

8. Additional Robustness Checks

We explore the substantive robustness of our main results to scale properties of the causal forest treatment effects, to "bootstrapping" individual-level errors, and in relation to other causal modeling approaches. Because these are ancillary analyses, we summarize the main findings here and provide full results in the Electronic Companion.

8.1. Robustness: Binary Analyses: Dichotomized Forest, RDIT, and GAM

We relax TCF's inverval-scaling assumptions through a series of binary analyses; "LogitForest" dichotomizes individual-level treatment effects via median split, whereas "LogitLinear" and "LogitGAM" relate whether the user engaged in each activity to covariates, respectively, linearly and with GAM fitting out-of-sample contours for *Age* and *Cohort*. Results reveal strong convergence for not only sign and significance but also *magnitude*; as per Table EC.9, all derived "treatment effects" are very similar, never deviating by more than 20% across analyses. Nontrivial differences in these models' underlying estimation methods—parametric, semiparametric, or nonparametric—lend credence to the notion that TCF reliably captures individual-level treatment effects.

8.2. Robustness: Heterogeneity: Bootstrapping Individual-Level TCF Errors

Our analyses are based on mean treatment effects ($\hat{\tau}_i$) and hold aside estimation imprecision, a quantity provided by TCF. Therefore, we replicated all regression-based analyses (LM, RLM, GAM) by bootstrapping 100 times, plotting their densities and summarizing comparative statistics. Results for the "causal" effects in the linear model (LM) with interactions appear in Table EC.10 and kernel density plots for the first week in Figure EC.11. These suggest that there is some degree of variation in estimated effects because of bootstrapping, but the resulting densities are always bounded away from zero, and, moreover, that the updated standard errors (including both sampling and bootstrapping errors) never alter the substantive conclusions.

8.3. Robustness: Heterogeneity with Other Causal Models

Given that the various modeling frameworks above provide convergent evidence on the direction and significance of treatment effects (see, e.g., Section 4.3 and Figure 7), one might presume this for covariate effects as well. But this is only partially the case, with some differing in sign and significance from the bulk of the others; all such results appear, for the first week after treatment, in Table EC.12. Generally speaking, TCF-based regression results (LM, RLM, and GAM) agree with the "modal" model; it is always in the majority in terms of the sign and significance of each of the covariate effects for each of the three DVs. Inconsistencies may arise in part because of many of these models specifying the exact functional form through which covariates enter. That is, in the language of hierarchical models, "Level II" has no error and relies on a linear-additive framework through which the treatment and covariate effects are jointly estimated. By contrast, TCF extracts individual-level treatment effects and then allows the analyst to use a variety of methods—linear (LM), robust (RLM), or semi-parametrically flexible (GLM)—through which to assess covariate effects.

Taken together, the robustness checks, placebo analyses, and simulation studies (as described in Appendix B.1) suggest that the standard TCI assumptions hold and that the proposed method—TCF—is able to recover individual-level treatment and covariate effects in both synthetic and, most critically, our real data.

9. 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 because of such invasions, especially so identity theft. Although much research has focused on measuring the effects of data breaches of companies whose financial 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 they could readily avail of causal inference methods to achieve the above objectives.

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 that are 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 the customer finds valuable. It may be that, even among customers who react negatively to the breach, some will respond to very different reparative incentives, for example, with some preferring security services (as in the well-known Equifax breach) and others financial concessions (as in the data breach to Target; see Kude et al. 2017).

In our empirical analysis, we find that 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 were initially

upset and reduced their activity accordingly, but, in relatively short order, we could see a hint of "life returns to normal." A variety of post-breach communications from the company suggested that the longer-term effect may well have been positive and so potentially echoes the work of Han et al. (2020) and Zhong and Schweidel (2020) regarding the eventual effects of negative shocks. Although we do not have sufficient data to either prove or rule this out, a possible consequence is that potential customers became aware of the existence of the website because of the breach, thereby increasing the likelihood of joining later on, when the heat of the breach dissipated, because of this increased awareness.

Importantly for privacy mitigation, we document that users who had a priori higher preferences for privacy were the last and least to react in protecting themselves. This might hint to the unclear messages of what could have been done to further protect oneself following the data breach. Again, whether this generalizes to other breaches is an empirical question, but it may be that firms should attend to breaches vigorously and assist their customers in protecting themselves from further loss whenever possible. Policymakers, users of compromised 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 incentivize 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. Although 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.

Acknowledgments

Special thanks are due to Elizabeth Bruch, Nigel Melville, and A. Yeşim Orhun for their extensive feedback on earlier versions of this paper. The authors thank their colleagues at the partnering company who provided invaluable guidance and ensured data quality. Although the project has benefited from all of this support, any errors are solely the authors'.

Appendix A. Linear Model (with Interactions) Heterogeneity Estimates

	Deleted photos			Searches			Sent messages		
	Week 1	Week 2	Week 3	Week 1	Week 2	Week 3	Week 1	Week 2	Week 3
Constant	2.805***	−0.166***	−0.300***	−0.289***	−3.363***	−3.369***	−1.426***	−2.425***	−1.921***
	(0.011)	(0.001)	(0.001)	(0.005)	(0.012)	(0.011)	(0.006)	(0.012)	(0.005)
Age	2.085***	0.046***	−0.053***	−0.460***	−1.758***	−1.520***	−0.581***	−1.263***	−0.525***
	(0.102)	(0.006)	(0.005)	(0.047)	(0.111)	(0.103)	(0.055)	(0.115)	(0.045)
Cohort	−0.041	0.372***	0.345***	1.263***	−0.652***	0.333***	−1.086***	−0.177	0.796***
	(0.099)	(0.006)	(0.005)	(0.045)	(0.107)	(0.100)	(0.053)	(0.111)	(0.043)
Married	0.529***	0.007***	−0.016***	−0.103***	−0.348***	−0.294***	−0.160***	−0.303***	−0.123***
	(0.024)	(0.001)	(0.001)	(0.011)	(0.026)	(0.024)	(0.013)	(0.027)	(0.010)
Private	−1.154***	−0.027***	0.034***	0.157***	0.766***	0.545***	0.289***	0.403***	0.209***
	(0.032)	(0.002)	(0.002)	(0.015)	(0.035)	(0.032)	(0.017)	(0.036)	(0.014)
AgeSquare	−0.318***	−0.007***	0.007***	0.066***	0.268***	0.223***	0.071***	0.149***	0.067***
	(0.018)	(0.001)	(0.001)	(0.008)	(0.019)	(0.018)	(0.010)	(0.020)	(0.008)
CohortSquare	0.498***	−0.161***	−0.198***	−0.738***	−0.317***	−0.624***	0.518***	−0.020	−0.467***
	(0.041)	(0.002)	(0.002)	(0.019)	(0.044)	(0.041)	(0.022)	(0.046)	(0.018)
Age:Married	0.242***	0.003	−0.007***	−0.053***	−0.151***	−0.119***	−0.084***	−0.146***	−0.065***
	(0.032)	(0.002)	(0.002)	(0.015)	(0.035)	(0.032)	(0.017)	(0.036)	(0.014)
Age:Cohort	−0.044	−0.008***	−0.002	0.029*	0.069*	0.067*	0.123***	0.143***	0.075***
	(0.029)	(0.002)	(0.001)	(0.013)	(0.031)	(0.029)	(0.016)	(0.032)	(0.013)
Age:Private	0.007	0.002	0.002	0.005	−0.095*	−0.067	0.013	−0.066	−0.010
	(0.043)	(0.002)	(0.002)	(0.020)	(0.046)	(0.043)	(0.023)	(0.048)	(0.019)
Cohort:Married	0.017	−0.030***	−0.012***	0.053**	0.030	0.032	0.197***	0.197***	0.117***
	(0.043)	(0.002)	(0.002)	(0.020)	(0.046)	(0.043)	(0.023)	(0.048)	(0.019)
Married:Private	−0.385***	0.002	0.016***	0.120***	0.312***	0.319***	0.141***	0.498***	0.186***
	(0.071)	(0.004)	(0.004)	(0.032)	(0.076)	(0.072)	(0.038)	(0.079)	(0.031)
Cohort:Private	−0.215***	0.052***	0.037***	−0.034	0.199**	0.224***	−0.382***	−0.409***	−0.101***
	(0.056)	(0.003)	(0.003)	(0.026)	(0.060)	(0.057)	(0.030)	(0.063)	(0.024)
Observations	49,993	49,993	49,993	49,993	49,993	49,993	49,993	49,993	49,993
Resid. StdErr (df = 49,980)	0.012	0.024	0.021	0.011	0.029	0.036	0.010	0.004	0.003
F Statistic (df = 12; 49,980)	129.9***	73.3***	79.0***	625.8***	408.4***	168.7***	939.1***	197.3***	4,070.2***

Notes. Age and Cohort are for Interquartile, not unit, changes; $IQ_A = 15$; $IQ_C = 11$. All covariates mean-centered, so marginal effects at centroid are given by linear terms

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Appendix B. Placebo Tests and Simulation Studies

Causal inference methodology always enacts identification assumptions. In the example of an exogenous shock such as the one analyzed here, a possible concern might be that assignment to cohorts (defined as the time of joining the site) is confounded with the timing or effects of the data breach itself. Although there is no apparent reason why this should be so, the construction of "Temporal Causal Inference" offers a possible remedy; because almost all cohorts are used both as treatments and controls for various time periods, possible differences in breach reactions are accounted for. These differences are discussed in Section 4.

Another assumption that should be scrutinized is "parallel trends," which also underlies Diff-in-Diff methodology. Although impossible to test directly, we find that the construction of the control and treatment groups resulted in indistinguishable time trends prior to the treatment (as per EC.1.2). Moreover, we used TCF to account for whatever possible differences may arise, down to the individual level.

B.1. Placebo Test

To further validate this assumption, we ran placebo tests prior to the announcement of the data breach, the general idea being that the method should not infer "treatment effects" when there were none. Specifically, we recreated the control and treatment groups via Temporal Causal Inference, with fake "treatment" (i.e., time where no data breach occurred) five weeks prior to the data breach, and then reran TCF with these data sets on all three types of activities.[26] The analyses showed nonsignificant or negligible effects of the treatment, as expected. Results appear in Figure B.1. The nature of these analyses allowed us to make further use of the placebo settings in a simulation study, as described next.

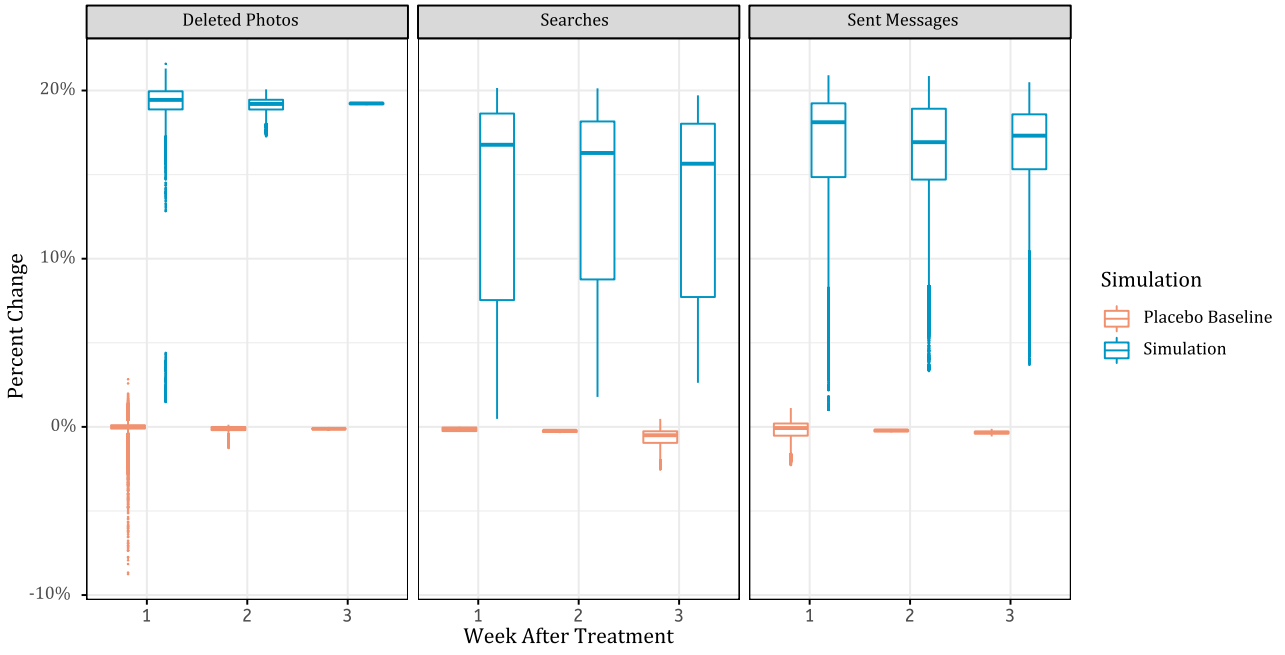
B.2. Placebo-Based (Real Data) Simulation

Because of the nature of the data set, we wanted to verify that, even with the noise and heterogeneity in site activities, it is possible to recover predefined treatment effects accurately. To do so, we used the placebo setting—pretending that the announcement happened five weeks prior to the data breach—and then created a "fake" treatment effect with an average size of 20%. Temporal Causal Forests methodology was able to recover the effects accurately; these results appear in Figure B.1.

B.3. Synthetic Simulation

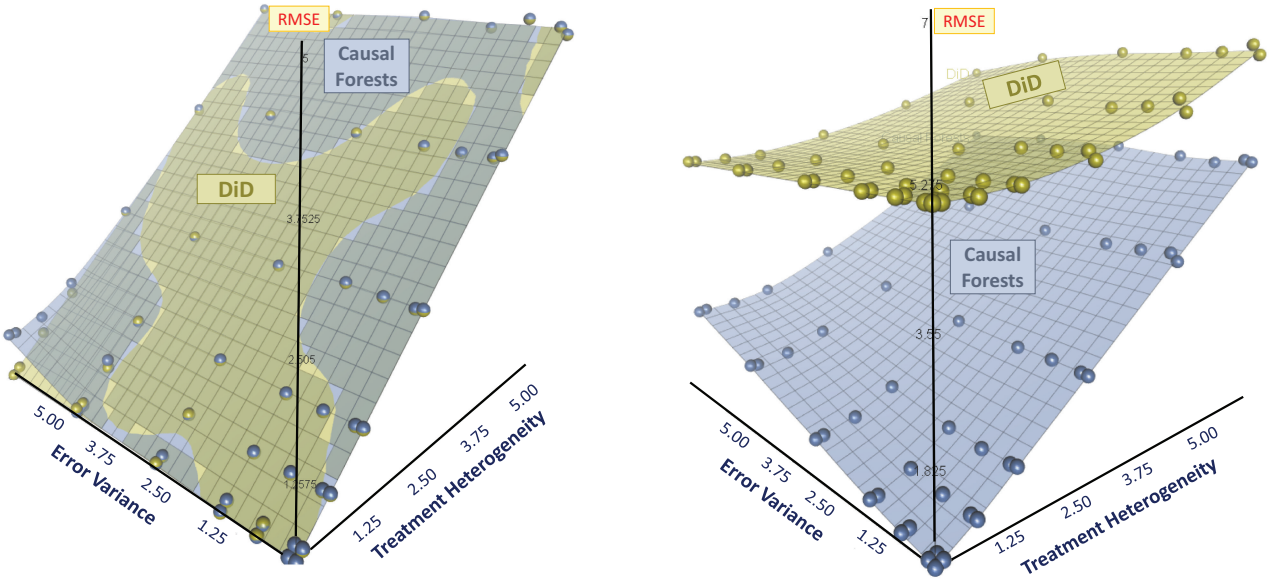
The second simulation is entirely synthetic and included data generated using a Diff-in-Diff model, with prespecified error around the covariates and around the treatment effect. The dependent variable in all synthetic simulations is log(number of activities) to mitigate both heteroscedasticity concerns and the confounding of location and scale intrinsic to binary data. We compared both (1) Temporal Causal Forests and (2) the model that generated the data, Diff-in-Diff (DID). Somewhat surprisingly, although DID served as the data-generating process, TCF recovers the effects very well in all prespecified settings (Figure B.2, left panel). Moreover, when the individual treatment was correlated with the individual fixed effect, via modeling it to be $\tau_i \sim N(\mu_i, \sigma_\tau)$, where μ_i is the individual fixed effect and σ_τ is the variance of the treatment in the population, Temporal Causal Forests outperformed DID (i.e., had

Figure B.1. (Color online) Results of Simulation Based on Placebo: Fake Treatment of 20% Increase in Likelihood of Being Active, Added to the Placebo Data



Notes. Box plots show the median and 25th to 75th percentiles of percent change. Placebo baseline in red (left of each week) and simulation in blue (right of each week)

Figure B.2. (Color online) Results of Simulation Studies: RMSE of Diff-in-Diff and Temporal Causal Forests as a Function of Treatment Heterogeneity and Error Variance



Note. Left panel shows a fit to the linear model that generated the data, right panel for a model where the treatment effect is correlated with the individual fixed effects.

lower RMSE) in all pre-defined variance settings (Figure B.2, right panel).

Endnotes

[1] Privacyrights.org lists measures through 2018. Information on following years was collected independently. The most recent "major" data breach at time of writing was Twitter in 2022, where hackers exposed 235 million users' information.

[2] The terms "customer" and "user" are used interchangeably throughout.

[3] Our analyses do not make use of any "hacked" data; rather, the data we use were provided directly by the firm and conform to a nondisclosure agreement. Because of confidentiality, we name neither the focal website nor the precise time range of the announcement, similarly omitting potentially identifying technical details regarding the nature of the announcement.

[4] "Three-quarters Facebook users as active or more since privacy scandal: Reuters/Ipsos poll", May 2018: www.reuters.com/article/us-facebook-privacy-poll/idUSKBN1I7081.

[5] The nature of our data resisted attempts to estimate this class of models, as discussed in Section 2.1.

[6] We also verified this assumption, to the extent possible, with detailed analyses of media publication dates and Google Trends around the name of the website, which spiked less than a day after the announcement.

[7] Although the number of photos users placed on the site was not observable, uploading at least one was required upon joining.

[8] Owing to the nature of purchasing blocks of credit "noncontractually" and then using individual credits for messaging when needed, churn (Fader and Hardie 2009, Lemmens and Gupta 2020) is neither observed nor well-defined. Therefore, our analyses will be based on observable engagement on the website—messaging, searching, and deletion of photos—although churn would be an appropriate metric in a contractual setting where (non-)renewals are observed. Moreover, although users could have been asked for their profile to be deleted from the website, it was revealed during the breach itself that deletion is not sufficient for most users to protect their data from hackers. The ability to delete profiles also changed multiple times during our data window in both the pricing and the existence of this option. Therefore, deletion of photos was the only observable measure users could take to protect their information from being revealed to both anyone who browses the website and potential future hackers.

[9] The website is mostly heterosexually oriented; only about 0.1% of the sample sought same-sex affairs. This option did not include marital status, and such users are coded as single for the purposes of analysis. To verify both honesty in reporting marital status and to assess the possibility of "open relationships," 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, before the announcement of the breach. Results show a very similar percentage of married users in the survey and on the website. Moreover, the survey results largely rule out possible "open relationships" or other forms of socially acceptable affair-seeking among attached men.

[10] Note that, because the user can change his or her profile (be it because of real changes in the user's life, changes to the user's privacy preferences, or other reasons), the covariates included in the analysis are potentially time varying. To maintain consistency in 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.

[11] In practice, in order to have similar numbers of users, we take as control group the last five cohorts that joined prior to the treated group. Reported results are robust to numbers of cohorts larger than three.

[12] We note that there is a mild, nonsignificant increase in the percentage of active users as cohorts join the website. As explained in

Electronic Companion EC.1.2, this does not challenge the focal identification methodology.

[13] At the risk of oversimplification, in "Causal Inference," all of the above splitting settings are chosen to ensure that there will be similar propensity to be treated, and the predicted pseudo-outcome for each observation will be similar between the control and treatment units within each leaf (Guo et al. 2021). See Athey and Wager (2019), Section 2.2, for further details on the splitting rule.

[14] Because using the $\{\hat{\tau}_i\}$ values in further analyses—for example, as dependent variables—does not directly incorporate their estimation uncertainty, such results must be interpreted with care (see Semenova and Chernozhukov 2021 for additional details). To that end, we accommodate this uncertainty for our focal analyses via bootstrapping, using individual-level standard error estimates, in Section 8.2.

[15] Results were robust to other specifications of a relatively large (>2,000) number of trees.

[16] Causal Forests are suitable for estimation of treatment effects whether the covariates and the outcomes are binary or not and whether they are time series or not, so long as the Causal Inference assumptions, described in Section EC.1, hold (Davis and Nielsen 2020, O'Neill 2021).

[17] In addition to $\hat{\tau}_i$, the method can also be used to export "doubly robust" scores, which yield weekly treatment effects for DVs of interest. We present these results as robustness checks in Electronic Companion EC.3.

[18] Reported results are robust to running the method on multiple outcomes simultaneously. However, when running on multiple outcomes, individual variances are unavailable because of extreme computational cost. In the results below, we use those from multiple estimates: one for each type of activity and for each week after the treatment.

[19] Viewing photos set to "public" requires registration, which can be done *gratis*. That is, anyone interested in seeing photos of other users can join for that purpose and access them. Setting images to "private," on the other hand, means that the private user's photos are shared only with those users who share their photos back, that is, in a reciprocal exchange of images.

[20] In lieu of tabular results, which are numerous and available in Appendix A, we present analyses for outcomes of interest—that is, the effects of the breach and covariates associated with them—visually via mean effects and associated (95%) confidence intervals. To preserve the meaningful intrinsic scale of the nonbinary covariates, we have not standardized. Instead, to ensure coefficient comparability across covariates, Figure 7 depicts not unit covariate changes but interquartile ranges, which were 15 for *Age* and 11 for *Cohort*.

[21] Even though we cannot determine exactly how many photos a user had at the time of the breach, prior deletion is a reasonably good indicator of how many photos users made available. Therefore, because past behavior (X_{it}) was used in the estimation of the treatment effect, we control for the number of photos via proxy information available in the data set.

[22] The choice to designate photos as private is the clearest available signal for privacy for this website. However, we note that this may be driven by multiple behavioral tendencies. For example, it may signal higher intrinsic/extrinsic privacy preferences (Lin 2022) or the feeling that sharing photos publicly would hurt their chances for a match. Whatever the root cause for the preference may be, users who, prior to the breach, chose to have private photos should have been more likely to delete their photos soon after the breach, owing to the increased scrutiny it generated.

[23] Run using the LM, RLM, and MGCV packages in R.

[24] These two forms of "multiple" forest—both run with the `grf` package in R—do not at present allow computation of standard errors; because their results are substantively highly similar to those from individual forests, which also provide greater flexibility, we do not pursue these further.

[25] Including higher-order terms and interactions did not produce superior fit or substantive differences. Note that including fixed effects eliminates the need for the treatment indicator Tr_i because it is user-specific and therefore correlates perfectly with the individual fixed effects. In order to balance the number of observations for each tenure (in weeks) on the website, the data are restricted to the eight-week window prior to the treatment. Note that this reduction still allows us to gather insights on the full trajectory because different users are in different membership age (tenure) categories at the time of the treatment. All reported results are robust to using the full timeline of each user, which by construction varies between cohorts. Because we observe three weeks after the treatment (announcement of the data breach) and because 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 post-announcement data.

[26] The last few cohorts needed to be removed for the placebo tests because they had too short a time on the site before the fake treatment took place.

References

Ablon L, Heaton P, Lavery DC, Romanosky S (2016) *Consumer Attitudes Toward Data Breach Notifications and Loss of Personal Information* (Rand Corporation, Santa Monica, CA).

Acquisti A, Brandimarte L, Loewenstein G (2020) Secrets and likes: the drive for privacy and the difficulty of achieving it in the digital age. *J. Consum. Psychol.* 30(4):736–758.

Acquisti A, Varian HR (2005) Conditioning prices on purchase history. *Marketing Sci.* 24(3):367–381.

Amir E, Levi S, Livne T (2018) Do firms underreport information on cyber-attacks? Evidence from capital markets. *Rev. Accounting Stud.* 23(3):1177–1206.

Athey S, Catalini C, Tucker C (2017) The digital privacy paradox: Small money, small costs, small talk. Working Paper Series, Technical report, National Bureau of Economic Research.

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *Ann. Statist.* 47(2):1148–1178.

Athey S, Wager S (2019) Estimating treatment effects with causal forests: An application. *Observational Stud.* 5(2):37–51.

Bachmann P, Meierer M, Näf J (2021) The role of time-varying contextual factors in latent attrition models for customer base analysis. *Marketing Sci.* 40(4):783–809.

Brandimarte L, Acquisti A, Loewenstein G (2013) Misplaced confidences: Privacy and the control paradox. *Soc. Psychol. Personal. Sci.* 4(3):340–347.

Calonico S, Cattaneo MD, Farrell MH, Titiunik R (2017) Rdrobust: Software for regression-discontinuity designs. *Stata J.* 17(2):372–404.

Calonico S, Cattaneo MD, Farrell MH, Titiunik R (2019) Regression discontinuity designs using covariates. *Rev. Econ. Stat.* 101(3):442–451.

Cattaneo MD, Titiunik R (2021) Regression discontinuity designs. *arXiv preprint arXiv:2108.09400*.

Chatterjee S, Gao X, Sarkar S, Uzmanoglu C (2019) Reacting to the scope of a data breach: The differential role of fear and anger. *J. Bus. Res.* 101:183–193.

Chen HS, Jai TM (2021) Trust fall: Data breach perceptions from loyalty and non-loyalty customers. *Service Indust J.* 41(13–14): 947–963.

Cleeren K, Dekimpe MG, van Heerde HJ (2017) Marketing research on product-harm crises: A review, managerial implications, and an agenda for future research. *J. Acad. Marketing Sci.* 45(5):593–615.

Davis RA, Nielsen MS (2020) Modeling of time series using random forests: Theoretical developments. *Electron. J. Stat.* 14(2):3644–3671.

Fader PS, Hardie BG (2009) Probability models for customer-base analysis. *J. Interactive Marketing* 23(1):61–69.

Goldfarb A, Tucker C (2012) Shifts in privacy concerns. *Amer. Econom. Rev.* 102(3):349–353.

Goldfarb A, Tucker C, Wang Y (2022) Conducting research in marketing with quasi-experiments. *J. Marketing* 86(3):1–20.

Gopalakrishnan A, Bradlow ET, Fader PS (2017) A cross-cohort changepoint model for customer-base analysis. *Marketing Sci.* 36(2): 195–213.

Granger CW (1980) Testing for causality: A personal viewpoint. *J. Econom. Dynam. Control* 2:329–352.

Guo T, Sriram S, Manchanda P (2021) The effect of information disclosure on industry payments to physicians. *J. Marketing Res.* 58(1):115–140.

Gwebu KL, Wang J, Wang L (2018) The role of corporate reputation and crisis response strategies in data breach management. *J. Management Inform. Systems* 35(2):683–714.

Han JA, Feit EM, Srinivasan S (2020) Can negative buzz increase awareness and purchase intent? *Marketing Lett.* 31(1):89–104.

Hastie T, Tibshirani R (1990) *Generalized Additive Models* (Chapman and Hall/CRC, London).

Janakiraman R, Lim JH, Rishika R (2018) The effect of a data breach announcement on customer behavior: Evidence from a multichannel retailer. *J. Marketing* 82(2):85–105.

Kim S, Lee C, Gupta S (2020) Bayesian synthetic control methods. *J. Marketing Res.* 57(5):831–852.

Knaus MC, Lechner M, Strittmatter A (2021) Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *Econom. J.* 24(1):134–161.

Kude T, Hoehle H, Sykes TA (2017) Big data breaches and customer compensation strategies: Personality traits and social influence as antecedents of perceived compensation. *Internat. J. Oper. Prod. Management* 37(1):56–74.

Lemmens A, Gupta S (2020) Managing churn to maximize profits. *Marketing Sci.* 39(5):956–973.

Li G (1985) Robust regression. Hoaglin DC, Mosteller F, Tukey JW, eds. *Exploring Data Tables, Trends, and Shapes* (Wiley, New York), 281–343.

Lin T (2022) Valuing intrinsic and instrumental preferences for privacy. *Marketing Sci.* 41(4):663–681.

Madden M, Rainie L (2015) *Americans' Attitudes About Privacy, Security and Surveillance* (Pew Research Center).

Martin KD, Borah A, Palmatier RW (2017) Data privacy: Effects on customer and firm performance. *J. Marketing* 81(1):36–58.

Massey FJ Jr (1951) The kolmogorov-smirnov test for goodness of fit. *J. Amer. Statist. Assoc.* 46(253):68–78.

O'Neill E (2021) Essays on Tree-based Methods for Prediction and Causal Inference. PhD thesis, University of Cambridge.

Romanosky S, Hoffman D, Acquisti A (2014) Empirical analysis of data breach litigation. *J. Empir. Leg. Stud.* 11(1):74–104.

Romanosky S, Telang R, Acquisti A (2011) Do data breach disclosure laws reduce identity theft? *J. Policy Anal. Management* 30(2): 256–286.

Rosati P, Cummins M, Deeney P, Gogolin F, van der Werff L, Lynn T (2017) The effect of data breach announcements beyond the stock price: Empirical evidence on market activity. *Int. Rev. Financ. Anal.* 49:146–154.

Rubin DB (2005) Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100(469):322–331.

Semenova V, Chernozhukov V (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *Econom. J.* 24(2):264–289.

Stan Development Team (2022) RStan: the R interface to Stan. URL https://mc-stan.org/, r package version 2.21.5.

Taylor CR (2004) Consumer privacy and the market for customer information. *RAND J. Econ.* 35(4):631–651.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113(523):1228–1242.

Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* 15(1):1625–1651.

Wang P, D'Cruze H, Wood D (2019) Economic costs and impacts of business data breaches. *Issues Inform. Systems* 20(2):162–171.

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

Zhong N, Schweidel DA (2020) Capturing changes in social media content: A multiplelatent changepoint topic model. *Marketing Sci.* 39(4):827–846.

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