Appendix EC.1: Causal Inference Assumptions in Temporal Causal Inference

Our causal inference mechanism builds on the "potential outcome framework", a term attributed to Neyman (Rubin 2005). According to the potential outcome framework, any causal inference problem relies on two quantities: $Y_i^{(0)}$ and $Y_i^{(1)}$, the outcome of unit i without and with receiving the treatment $W_i \in \{0,1\}$, respectively. The typical measure of the treatment effect is $Y_i^{(1)} - Y_i^{(0)}$. A fundamental problem of causal inference is that, in any experiment – may it involve randomly assignment or not – the researcher cannot ever observe both $Y_i^{(0)}$and $Y_i^{(1)}$, as the user is either exposed to the treatment, or not. Therefore, the researcher should estimate the missing quantity – referred to as the potential/counterfactual outcome – either $\hat{Y}_i^{(0)}$or $\hat{Y}_i^{(1)}$, and use the counterfactual outcome to infer the causal effect. In order to do that, we should also estimate the probability to be treated, which we denote $\hat{w}_i \in (0,1)$ (the restriction of not being 0 or 1 is explained in Section EC.1.3).

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

EC.1.1. SUTVA (Stable Unit Treatment Value Assumption)

The SUTVA assumption (Rubin 1980) comprises two conditions:

a. No interference between units (Cox 1958). Neither $Y_i^{(1)}$ nor $Y_i^{(0)}$ is affected by the treatment assignment any other unit received: $Y_i^{(W_i)} \perp Y_j^{(W_j)}$ for any two users $i, j \in 1, \ldots N$. In the case of measuring the effects of an exogenous shock, this assumption means that if one user/customer is exposed to the shock, this should not affect the outcome of

any other user. By design of TCI, assignment to treatment or control groups is based on the time of joining the website. Since the outcomes do not occur at the same real-world-time, no two users who were in the *different* groups – control and treatment – are affected by the other's treatment or lack thereof. That is, for every set $\{i, j\}$, $Y_i^{(1)} \perp Y_j^{(0)}$. On the other hand, behavior following the data breach of one treated user, might affect another user, that is, it is possible that $Y_i^{(1)} \perp Y_j^{(1)}$ for some *treated* users i, j. This is due to the nature of the website: a match-making website, where one user's change in behavior (e.g., stopping using the website) might affect another user's behavior (e.g., enjoys the website more, now that there is less "competition" on the website). In addition, due to network effects, it might be the case that $Y_i^{(0)} \perp Y_j^{(0)}$. Since our objective is to measure the effect of the data breach, and in general – the effect of other exogenous shocks – this possible dependency is not a concern; rather, it **should** be part of the estimation of the treatment effect.

As another possible interference due to network effect, it is possible that the activities of users in the control group affected users in the treated group (who joined later by construction of TCI). However, this should not undermine the ability to identify the treatment effect of the data breach. The website was in stable condition in terms of growth – it was a mature website and we can see evidence for this in the comparison of the control and treatment groups provided below. Moreover, if men in the treated group are affected by the behavior of those who preceded their joining and are in the control group, this effect is independent of the treatment assignment – since no one could anticipate it. On a matter of negation, let's assume that there is interference between units – for example, that users in the control group, who preceded the joining of those in the treatment group, affected the behaviors of those in the treated group. Similar claim can therefore be made regarding to the control group users relative to those who preceded *their* joining, due to the stability of the website. The control group's activities were therefore similarly affected by those who preceded their joining. Since we compare control group's activities to those of the treated units, such similar effects, even if indeed exist, would have been accounted for by the mere comparison of control and treatment units.

In other scenarios where the website or service are not stable, and there is interaction between units or other network effects, it is imperative to control for possible changes in users' trajectories that are not merely the treatment effect. In order to do this, the researcher can account for shifts in behavior through time or to remove placebo effects

if those are found. Even though we do not see such effects on the average user, we nevertheless complement our identification with Causal Forests, to match users' trajectories above and beyond their tenure on the website.

Of note, Temporal Causal Inference is not directly related to staggered differences in differences method. Whereas Staggered differences in differences is aiding in estimating treatment effects when the treatment happened at different times to different entities, in our settings, everyone were treated at the same time, but each cohort was at different phase (number of weeks) since joining the website. Nevertheless, it is imperative to assure all causal inference assumptions are held, to avoid biases such as the ones described in Baker et al. (2021).

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

As noted by Rubin (2005), under randomized controlled trials (RCT), there is no need for any assumptions other than SUTVA. However, we cannot avail of an RCT, and therefore we need to assure our models have properties that are "given" in RCTs. We therefore

proceed with showing that TCI allows for recovery of the treatment effect of an exogenous shock, while holding the necessary following assumptions:

EC.1.2. Conditional Independence (or Ignorability) Assumption

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

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

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

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

The data breach, as an exogenous shock, affected all users. However, because some users joined the website earlier, they were affected by the breach at a later point in their lifetime on the website, for unobserved reasons that might be confounded with the treatment effect (e.g., if for some reason, people that are more active, selected to join in a specific month). Therefore, the construction of TCI might not overcome this. In order to assure that the treatment group has similar behavior to that of the control group, prior to the treatment, we test the parallel trend assumption using a Granger Test (Granger 1980)[27]. We also find that, across all treatment groups and across all activities, the control groups constructed using Temporal Causal Inference act as suitable predictors for the behavior of their respective treatment group. Lastly, we conducted a Kolmogorov-Smirnov Test (Massey Jr. 1951) to verify that the control and treatment groups do not differ in their timeline prior to the treatment. Specifically, for each activity and for each treatment-control pair, we compute a cumulative sum of the average percent of active users prior to the treatment and divide it by the groups' maximum cumulative sum; this effectively creates a CDF-like timeline, as shown in Figure EC.1. We then conduct a Kolmogorov-Smirnov Test to test whether the timelines differ between the control and treatment groups. We find that, throughout all activities and treatment groups, the trends are indistinguishable from those of the respective control groups. Tabular results for both the Granger Causality and Kolmogorov-Smirnov tests are available in EC.13.

[27] The Granger Causality Test is a common way to compare two time series. Despite "causality" in its title, it is well known as a test of predictability, or "temporal relatedness," of one time series to another, and should not be misconstrued as a test of causality. Nevertheless, we implemented a bi-directional test of predictability of the control group on the treatment group, to assess whether the time series of the control group predicts that of the respective treated group prior to the treatment. We found that for all treatment groups and for all activities, the time series were statistically indistinguishable from the respective control groups (the null hypothesis of no-prediction was rejected with $p < 0.001$ across all tests).

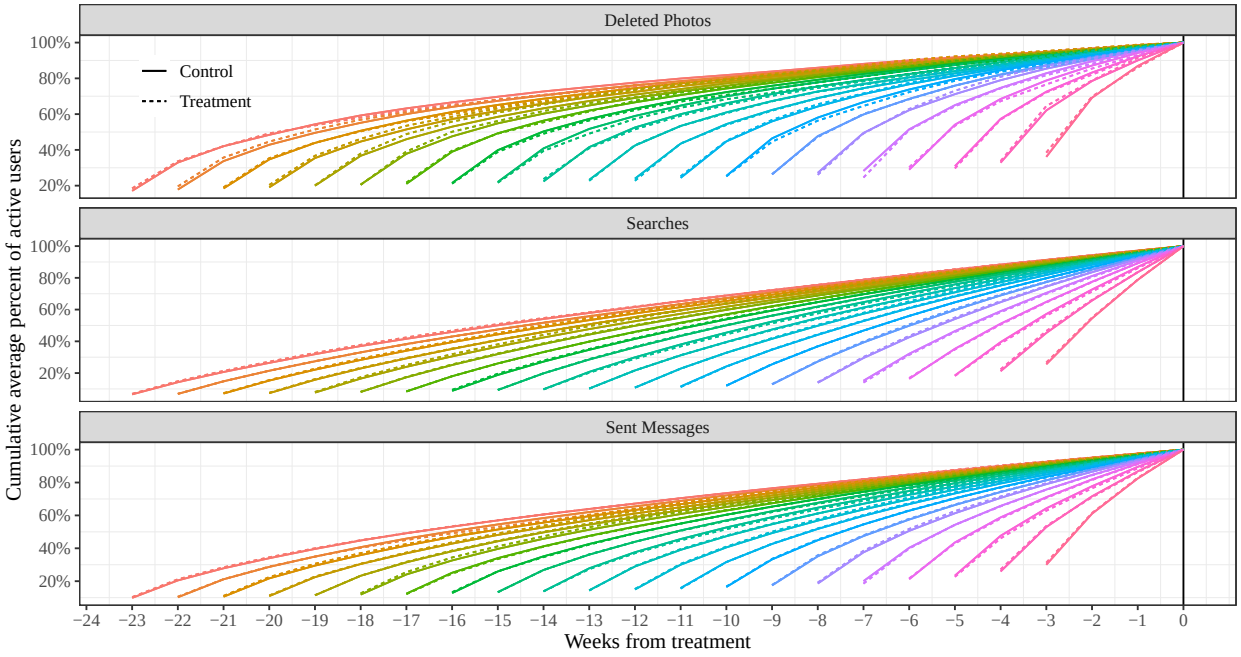


Figure EC.1 **Comparison of control and treatment groups: CDF-like comparison of treatment and control groups. Each line is the cumulative sum of the percent of active users for this group, divided by the maximum cumulative sum for this group (therefore always gets to 100%, and starting "higher" for cohorts with shorter timelines, due to this).**

Even though we established that the timelines are not statistically different from one another, and therefore there are parallel trends between the groups, in order to estimate individual treatment effects while also acknowledging that users might have varying timelines, we will add a second step to our causal inference method, one that will account for possible differences in the timelines of users prior to the treatment. In short, we will match users in the control and treatment groups, nonparametrically, using Temporal Causal Forests.

EC.1.3. Overlap Assumption (or "Common Support" Assumption)

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

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

In TCI, due to the nature of the exogenous shock, every user had some propensity to have been treated in any week during his period on the site; evidently, all users were treated. Nevertheless, we further tested this assumption by observing the estimated $\tilde{w}\left(x\right)$ – the propensity to be treated – estimated directly using Local Linear Forests (Friedberg et al. 2020) and as a by-product of Causal Forests.

EC.1.4. Exogeneity of Covariates Assumption

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

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

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

Appendix EC.2: Temporal Causal Forests

We introduce two changes to the original Causal Forest method. These changes are both internal and external to the estimation of the treatment effect, and were found, in a series of simulation studies we ran (described in Appendix B.1), to provide the best results in terms of RMSE and ability to recover heterogeneous treatment effects, both in synthetic data and on placebo tests on our dataset. These changes are:

1. In most cases, the use of the Causal Forest is to generate groups that are equivalent in both their propensity to be treated, and their expected outcome, given the covariates. We choose the parameters X_i to include the *time trend*. This leverages the traditional Causal Forests framework to group users based on their pattern of activities throughout time, resulting in groups within the control and treatment groups that are relatively homogeneous in respect to their *time trend*. In other words, TCF allows us to assess individual treatment effects by estimating a counterfactual time trend of their activities. Therefore, TCF allows treatment and control groups to be compared while verifying that the users have relatively similar time trends before the breach announcement (i.e., the treatment).

2. In order to improve Causal Forests, we also ran as robustness an analysis where estimate its "nuisance parameters" (to be explicated later) using Local Linear correction, via Local Linear Forests (Friedberg et al. 2020), further described below. However, in the main analyses, where we estimate all cohorts jointly, this estimation was not feasible due to the unequal length of timelines.

EC.2.1. Construction and Estimation – Temporal Causal Forests

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

a. Random Forests is a supervised machine learning method aimed at estimating a prediction $\hat{\mu}(x)$ for a vector of covariates $X_i = x$. The estimation can be seen as an "ensemble method", by taking the average of all regression/decision trees. Each decision tree is built by splitting the data into two leaves, in a greedy way, to minimize the Sum-of-Squared Error between the observed and predicted outcomes: consider a parent node P with n_P observations, $(X_{i1}, Y_{i1}), \ldots, (X_{in_P}, Y_{in_P})$. For each candidate pair of child nodes, $\{C_1, C_2\}$, let $\bar{Y}_1$, $\bar{Y}_2$ be the corresponding mean of Y in that leaf. The chosen pair of child nodes will be those that minimize:

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

Specifically, we present here "Honest Forests": for each tree $b \in \{1, \ldots, B\}$, where B is the number of trees in the forest, draw a subsample S_b, referred to as the training sample, in the size of half of the population (the size can be tuned). Grow the regression tree by recursively splitting so that the error function (to be defined for each problem separately) will be optimized. After training the forest for each user with set of covariates x not in S_b, make out-of-bag predictions on the response variable, $\hat{\mu}(x)$:

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

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

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

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

Therefore, the weights $\alpha_i(x)$ are higher for the observations that appear more often in the same leaf as x – and are thus "closer to it". Note that by construction, $\sum_{i=1}^{N} \alpha_i(x) = 1$, and $\alpha_i(x) \geq 0$.

c. Local Linear Forests. Whereas in Athey et al. (2019) the nuisance parameters, $\hat{w}(x)$ and $\hat{y}(x)$, were estimated via regression forests, Friedberg et al. (2020) demonstrate how to achieve better accuracy by estimating $\hat{w}(x)$ and $\hat{y}(x)$ using *Local Linear Forests*. These build on Generalized Random Forests, adding a "layer" of linear regression to exploit smoothness of the outcome, to correct for potential misalignment between a test point and its neighborhood, and potentially resolve instances of unbalanced data and noise. We therefore add a layer of Local Linear Forests as a robustness check when estimating the treatment effects for each cohort separately[28].

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

```
X = is(x_it >0)
Y = is(y_i >0)
W = is(i in treatment group)

if(each cohort is estimated separately)
   Y.hat = local_linear_forest(X,Y)
   W.hat = local_linear_forest(X,W)

else
   Y.hat = regression_forest(X,Y)
   W.hat = regression_forest(X,W)

tau.hat = causal_forest(X,Y,W,Y.hat,W.hat)
```

[28] At the time of analysis, Local Linear Forests in the GRF package does not support varying lengths of covariate vectors.

Appendix EC.3: Doubly Robust Treatment Effects Comparison

	Deleted Photos			Searches			Sent Messages		
	Week 1	Week 2	Week 3	Week 1	Week 2	Week 3	Week 1	Week 2	Week 3
Best Linear Projection	2.740	−0.215	−0.354	−0.492	−3.563	−3.739	−1.573	−2.832	−2.143
SD	(0.109)	(0.077)	(0.073)	(0.177)	(0.182)	(0.185)	(0.158)	(0.160)	(0.161)
Mean of $\hat{\tau}$	2.805	−0.166	−0.300	−0.289	−3.363	−3.369	−1.426	−2.425	−1.921
SD	(0.012)	(0.001)	(0.001)	(0.005)	(0.012)	(0.011)	(0.006)	(0.012)	(0.005)

Note: Doubly robust treatment effects arise via the intercept from a "best linear projection" using the identical model on which the overall treatment effects ($\hat{\tau}$) of Table EC.4 and Table EC.5 are based. We note that the *individual-level* doubly robust best linear projection scores were found to lie outside the relevant underlying scale for percentage deviations, and must be interpreted with caution. They are presented here as evidence of robustness of the *group-level* averages.

Appendix EC.4: Treatment Effects: Overall, LM, RLM, GAM

DV	Week	Model	Interactions	Mean	SE	Result
Photos	1	OVERALL	—	2.8045	0.0115	POS
Photos	1	LM	Yes	2.8045	0.0109	POS
Photos	1	RLM	Yes	2.5892	0.0106	POS
Photos	1	GAM	Yes	2.8045	0.0109	POS
Photos	2	OVERALL	—	-0.1657	0.0007	NEG
Photos	2	LM	Yes	-0.1657	0.0006	NEG
Photos	2	RLM	Yes	-0.1662	0.0005	NEG
Photos	2	GAM	Yes	-0.1657	0.0006	NEG
Photos	3	OVERALL	—	-0.2996	0.0007	NEG
Photos	3	LM	Yes	-0.2996	0.0005	NEG
Photos	3	RLM	Yes	-0.2874	0.0005	NEG
Photos	3	GAM	Yes	-0.2996	0.0005	NEG
Searches	1	OVERALL	—	-0.2886	0.0052	NEG
Searches	1	LM	Yes	-0.2886	0.0050	NEG
Searches	1	RLM	Yes	-0.1793	0.0048	NEG
Searches	1	GAM	Yes	-0.2886	0.0049	NEG
Searches	2	OVERALL	—	-3.3631	0.0124	NEG
Searches	2	LM	Yes	-3.3631	0.0117	NEG
Searches	2	RLM	Yes	-3.0093	0.0109	NEG
Searches	2	GAM	Yes	-3.3631	0.0115	NEG
Searches	3	OVERALL	—	-3.3686	0.0115	NEG
Searches	3	LM	Yes	-3.3686	0.0110	NEG
Searches	3	RLM	Yes	-3.2106	0.0114	NEG
Searches	3	GAM	Yes	-3.3686	0.0109	NEG
Messages	1	OVERALL	—	-1.4257	0.0060	NEG
Messages	1	LM	Yes	-1.4257	0.0059	NEG
Messages	1	RLM	Yes	-1.2353	0.0046	NEG
Messages	1	GAM	Yes	-1.4257	0.0058	NEG
Messages	2	OVERALL	—	-2.4247	0.0123	NEG
Messages	2	LM	Yes	-2.4247	0.0121	NEG
Messages	2	RLM	Yes	-2.0956	0.0104	NEG
Messages	2	GAM	Yes	-2.4247	0.0120	NEG
Messages	3	OVERALL	—	-1.9214	0.0049	NEG
Messages	3	LM	Yes	-1.9214	0.0047	NEG
Messages	3	RLM	Yes	-1.8010	0.0041	NEG
Messages	3	GAM	Yes	-1.9214	0.0046	NEG

Appendix EC.5: Treatment Effects: Robustness to Alternative Forest Specifications

DV	Week	Model	Interactions	Mean	SE	Result
Photos	1	OVERALL	—	2.8045	0.0115	POS
Photos	1	MultiForest	—	2.7459	0.0093	POS
Photos	1	MultiMulti	—	2.8760	0.0149	POS
Photos	2	OVERALL	—	-0.1657	0.0007	NEG
Photos	2	MultiForest	—	-0.3112	0.0051	NEG
Photos	2	MultiMulti	—	-0.1088	0.0066	NEG
Photos	3	OVERALL	—	-0.2996	0.0007	NEG
Photos	3	MultiForest	—	-0.4844	0.0051	NEG
Photos	3	MultiMulti	—	-0.2870	0.0062	NEG
Searches	1	OVERALL	—	-0.2886	0.0052	NEG
Searches	1	MultiForest	—	-0.3054	0.0171	NEG
Searches	1	MultiMulti	—	-0.1422	0.0149	NEG
Searches	2	OVERALL	—	-3.3631	0.0124	NEG
Searches	2	MultiForest	—	-3.6260	0.0237	NEG
Searches	2	MultiMulti	—	-3.2408	0.0199	NEG
Searches	3	OVERALL	—	-3.3686	0.0115	NEG
Searches	3	MultiForest	—	-3.6909	0.0249	NEG
Searches	3	MultiMulti	—	-3.2569	0.0201	NEG
Messages	1	OVERALL	—	-1.4257	0.0060	NEG
Messages	1	MultiForest	—	-1.4205	0.0171	NEG
Messages	1	MultiMulti	—	-1.2707	0.0145	NEG
Messages	2	OVERALL	—	-2.4247	0.0123	NEG
Messages	2	MultiForest	—	-2.5726	0.0200	NEG
Messages	2	MultiMulti	—	-2.2914	0.0169	NEG
Messages	3	OVERALL	—	-1.9214	0.0049	NEG
Messages	3	MultiForest	—	-2.1131	0.0204	NEG
Messages	3	MultiMulti	—	-1.7200	0.0156	NEG

Appendix EC.6: Alternative Causal Models: DiD, RDIT, GSynth, BSCM

In this section, we further detail two of the alternative causal models, and provide detailed results. Key details and results appear in the paper proper.

- **RDIT Model Specifications:** We conducted an RDIT analysis on each of the three focal DVs using a linear or logit link, using both dedicated packages (here, `rdrobust` and `rdlocrand` for parametric and local nonparametric models, respectively; Calonico et al. 2015) and custom code. For each DV, many parametric models are fitted, according to combinations of link (linear, binary), inclusion of individual-level covariates (none vs. {Age, Married, Private} linearly vs. treatment effect interactions), week-level fixed effects (included vs. not), and up to 7th-order polynomial time trends; that is, $84 = 2 \times 3 \times 2 \times 7$ in total[29]. For all three DVs and both link functions, we found that including covariates, their second-order interactions, weekly fixed effects, and 6th-order time trends provided the best results, based on parameter significance, predictive stability, and synthetic measures (e.g., AIC and BIC).

- **GSynth Model Specifications:** Owing to computational intensity, all analyses using Generalized Synthetic Control Method were conducted at the cohort-level, thus extracting average effects, and not heterogeneity in them. Moreover, GSynth requires at least 6 weeks prior to the treatment to construct a parallel trend between the control and treatment groups, so that results include only those cohorts on the website at least 6 weeks prior to the treatment; by contrast, for TCF we use all estimated cohorts. We conducted a model selection exercise, choosing among several specifications (adding covariates vs. not, adding cubic terms vs. not), and two estimation procedures – Matrix Completion (Athey et al. 2021) and Interactive Fixed Effects (Bai 2009) – presenting the best-fitting model in the pre-treatment periods. The best-fitting model was estimated using Matrix Completion, and is of the form:

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

where y_{jt} is the *proportion* of active users in cohort j at time t; $P_{jt} = 1$ if period t is post-treatment (or lack of treatment, for the control group) for cohort j, 0 otherwise; treatment indicator $Tr_j = 1$ iff cohort j is in the treatment group; μ_j is the fixed effect for cohort j and ϵ_{jt} denote zero-mean Gaussian (across cohorts) error terms.

[29] {Age, Married, Private} are mean-centered so that interactions are meaningful in models that include them.

EC.6.1. Results for Alternative Causal Models

Comparative results for all four alternative causal models – DiD, RDIT, GSynth, BSCM – appear in Table EC.6. Of note, RDIT provides treatment effect estimates for the first week post-announcement only, as it detects a "jump" associated with the breach. Very generally speaking, the results are broadly consistent across methods, always agreeing in sign, and nearly always in significance. If we consider each method *relative to* the others, several intriguing trends emerge. The least consistent method is DiD: despite selection across many candidate models, the magnitude of its treatment effects were exaggerated, sometimes greatly so, relative to the other methods, for example, being ten times greater for Photos in Weeks 2 and 3 and approximately double in Searches and Messages. BSCM was generally better, but produced estimates roughly 40% larger than the "central tendency" of the other methods. There was most commonality between GSynth and the proposed TCF-based method, despite substantial differences in their operationalizations, although GSynth produced a clearly anomalous result for Searches in Week 1 (roughly 1/4 of TCF and much lower than all the others).

All in all, 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) $\times$ 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 sub-section we examine robustness for treatment effect heterogeneity, as well as to individual-level errors.

DV	Week	Model	Interactions	Mean	SE	Result
Photos	1	OVERALL	—	2.8045	0.0115	POS
Photos	1	DiDLinear	—	1.6531	0.1076	POS
Photos	1	**RDITLinear**	Yes	2.1673	0.1389	POS
Photos	1	**RDITNon**	—	1.5758	0.1446	POS
Photos	1	GSynth	—	2.5939	0.1777	POS
Photos	1	BSCM	—	2.6436	0.0268	POS
Photos	2	OVERALL	—	-0.1657	0.0007	NEG
Photos	2	DiDLinear	—	-1.5843	0.1070	NEG
Photos	2	GSynth	—	-0.1685	0.2168	ns
Photos	2	BSCM	—	-0.3595	0.0284	NEG
Photos	3	OVERALL	—	-0.2996	0.0007	NEG
Photos	3	DiDLinear	—	-2.0372	0.1068	NEG

DV	Week	Model	Interactions	Mean	SE	Result
Photos	3	GSynth	—	-0.3229	0.2002	ns
Photos	3	BSCM	—	-0.4912	0.0285	NEG
Searches	1	OVERALL	—	-0.2886	0.0052	NEG
Searches	1	DiDLinear	—	-3.1017	0.1747	NEG
Searches	1	**RDITLinear**	Yes	-1.8517	0.2931	NEG
Searches	1	**RDITNon**	—	-3.6885	0.3280	NEG
Searches	1	GSynth	—	-0.0729	0.5641	ns
Searches	1	BSCM	—	-1.1702	0.0411	NEG
Searches	2	OVERALL	—	-3.3631	0.0124	NEG
Searches	2	DiDLinear	—	-6.8789	0.1774	NEG
Searches	2	GSynth	—	-3.1887	0.7514	NEG
Searches	2	BSCM	—	-4.4001	0.0452	NEG
Searches	3	OVERALL	—	-3.3686	0.0115	NEG
Searches	3	DiDLinear	—	-7.3812	0.1791	NEG
Searches	3	GSynth	—	-3.3606	0.7940	NEG
Searches	3	BSCM	—	-4.6913	0.0461	NEG
Messages	1	OVERALL	—	-1.4257	0.0060	NEG
Messages	1	DiDLinear	—	-4.8382	0.1736	NEG
Messages	1	**RDITLinear**	Yes	-3.4581	0.2511	NEG
Messages	1	**RDITNon**	—	-4.6798	0.2870	NEG
Messages	1	GSynth	—	-1.3240	0.5411	NEG
Messages	1	BSCM	—	-2.0517	0.0401	NEG
Messages	2	OVERALL	—	-2.4247	0.0123	NEG
Messages	2	DiDLinear	—	-6.5954	0.1755	NEG
Messages	2	GSynth	—	-2.4110	0.5910	NEG
Messages	2	BSCM	—	-3.2851	0.0423	NEG
Messages	3	OVERALL	—	-1.9214	0.0049	NEG
Messages	3	DiDLinear	—	-6.7898	0.1764	NEG
Messages	3	GSynth	—	-1.9569	0.6273	NEG
Messages	3	BSCM	—	-3.0588	0.0426	NEG

Appendix EC.7: Additional Robustness Checks

As summarized in Section 8 of the main text, we explore the substantive robustness of our main results relative to scale properties of the causal forest treatment effects; to "bootstrapping" individual-level errors; and in relation to extracted treatment effects of other causal modeling approaches.

EC.7.1. Robustness – Binary Analyses: Dichotomized Forest, RDIT, and GAM

Our analyses of the individual-level TCF treatment effects relies on their scale properties; even though these are based on nonparametric regularization, these may be affected in various ways by the method itself or artifacts in the data. We relax these by conducting a series of binary (logit) analyses with strongly differing assumptions. "LogitForest" takes the individual-level treatment effects and performs a median split, so that all covariates and their interactions are linearly related to the logit of that probability. "RDITLogit" runs the same regression discontinuity model as earlier, but using a logit link. Both "LogitLinear" and "LogitGAM" take the *original* activity data – whether the user engaged in each activity – and relate them to covariates in the usual linear-in-parameters way, with GAM fitting out-of-sample estimated contours for both *Age* and *Cohort*.

Results reveal strong convergence on not only sign and significance, but also in the *magnitude* of these effects. As per Table EC.9, all derived "treatment effects" are very similar, never deviating by more than 20% across all analyses. Given the sizable differences in these models' underlying estimation methods – some parametric, some semiparametric, some nonparametric – this lends some credence to the notion that TCF reliably captures individual-level treatment effects.

Having demonstrated that the various causal inference methods provide roughly comparable estimates – albeit with most showing some anomaly absent from TCF – we next turn to the robustness of the covariate effects.

EC.7.2. Robustness – Heterogeneity: Bootstrapping Individual-Level TCF Errors

Our analyses of the TCF-based treatment effects are based on their mean values, and hold aside the imprecision with which they are estimated, a quantity provided by TCF. We therefore replicated all analyses that apply regression methods (LM, RLM, GAM) to the TCF estimates by bootstrapping them 100 times, plotting their densities and summarizing their 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 results suggest that there is some degree of variation in estimated effects due to bootstrapping, but that their densities are always bounded away from zero; and, moreover that the updated standard errors – that is, including both sampling and bootstrapping error – do not more than double, never altering the substantive conclusion.

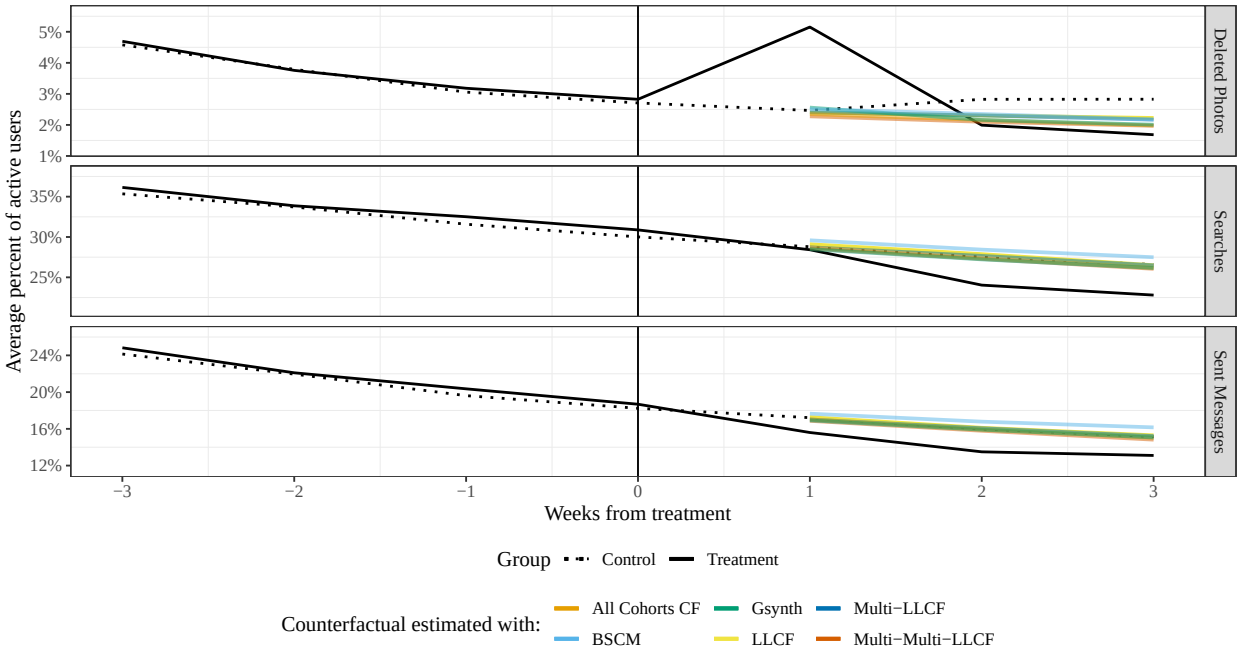
EC.7.3. Robustness – Heterogeneity with Other Causal Models

One might surmise that, given that the various modeling frameworks above provide convergent evidence on the direction and significance of treatment effects, they would do so in regard to covariate effects as well, in line with those reported in Section 4.3 and depicted in Figure 7. But this is only partially the case, with many covariate effects differing in sign and significance from the bulk of the others.

For conciseness, all such results appear, for the first week after treatment, in Table EC.12. Generally speaking, the results from the regression approaches (LM, RLM, and GAM) using the TCF estimates 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. By contrast, to take one particular example, the RDITLogit seems to disagree with the other models at least as much as it agrees, particularly for Messages. We believe that part of the inconsistency in estimates of covariate effects here is that many of these models *specify 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 semiparametrically flexible (GLM) – through which to assess covariate effects, thus metaphorically allowing for error in Level II because of individual-level heterogeneity in the extracted treatment 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 both in synthetic and, most critically, our real data.

Appendix EC.8: Treatment Effects: Visual Presentation of Robustness to Alternative Forest and Causal Specifications



Average percent of active users: Control (dashed) and Treatment (solid) groups (average across cohorts in each group), three weeks prior and post treatment, alongside the estimated counterfactual number of activities for the treated group, estimated with various other methods and specifications. Specifically: different Causal Inference Methods as described in 7.3, Bayesian Synthetic Control Method (BSCM), Generalized Synthetic Control Group (Gsynth); different Causal Forest Specifications, the main analysis (All Cohorts CF) where all cohorts are estimated jointly, but each week and each DV is estimated on its own, Causal Forests with Local Linear correction of the estimation of the nuisance parameters (LLCF), as described in Section 3.2.1 (note that LLCF analyses could have been made only on a single cohort at a time), LLCF estimated for multiple weeks (Multi-LLCF), and all weeks and all DVs at the same time (Multi-Multi-LLCF), as described in Section 6.2. As is apparent, these results are similar across-the-board. Numerical results appear in electronic companions EC.5 and EC.6.

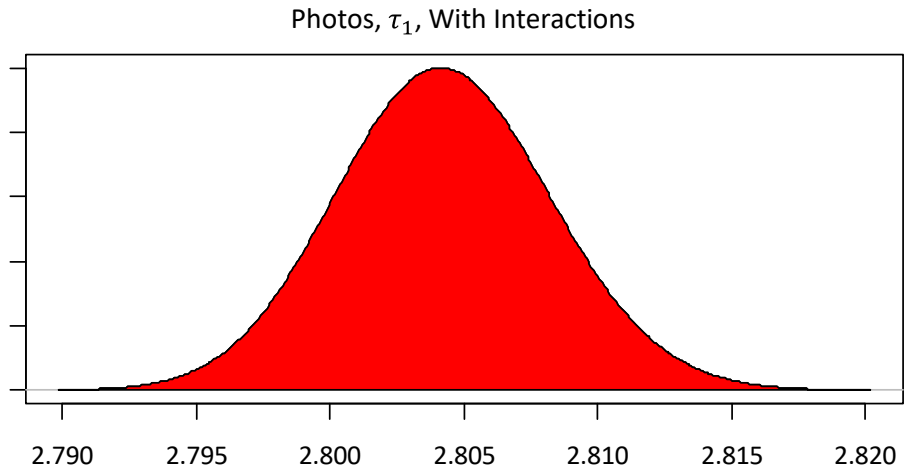
Appendix EC.9: Treatment Effects for Binary DV Specifications

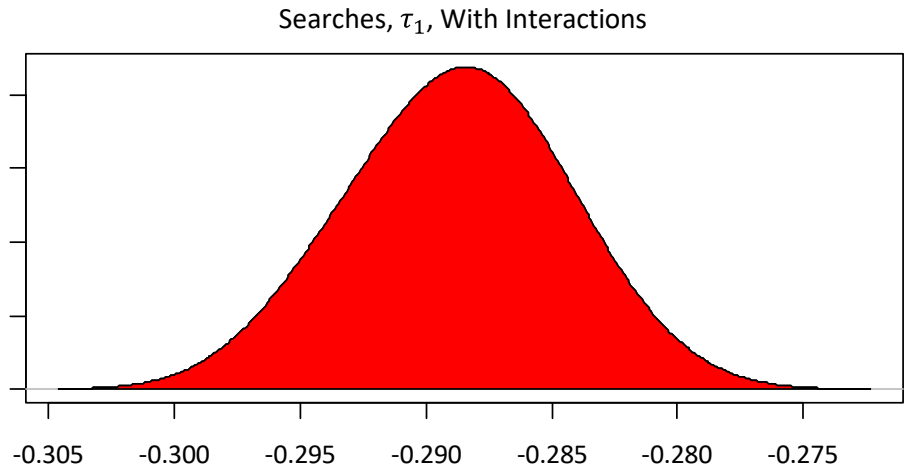
DV	Week	Model	Interactions	Mean	SE	Result
Photos	1	LogitForest	Yes	0.6430	0.0410	POS
Photos	1	**RDITLogit**	Yes	0.6993	0.0363	POS
Photos	1	LogitLinear	Yes	0.5848	0.0371	POS
Photos	1	LogitGAM	Yes	0.5000	0.0366	POS
Photos	2	LogitForest	Yes	-0.2580	0.0540	NEG
Photos	2	LogitLinear	Yes	-0.3197	0.0501	NEG
Photos	2	LogitGAM	Yes	-0.3930	0.0492	NEG
Photos	3	LogitForest	Yes	-0.3680	0.0590	NEG
Photos	3	LogitLinear	Yes	-0.4776	0.0557	NEG
Photos	3	LogitGAM	Yes	-0.5697	0.0553	NEG
Searches	1	LogitForest	Yes	-0.0770	0.0140	NEG
Searches	1	**RDITLogit**	Yes	-0.0954	0.0153	NEG
Searches	1	LogitLinear	Yes	-0.1251	0.0135	NEG
Searches	1	LogitGAM	Yes	-0.1099	0.0137	NEG
Searches	2	LogitForest	Yes	-0.2650	0.0160	NEG
Searches	2	LogitLinear	Yes	-0.3418	0.0157	NEG
Searches	2	LogitGAM	Yes	-0.3188	0.0160	NEG
Searches	3	LogitForest	Yes	-0.2880	0.0190	NEG
Searches	3	LogitLinear	Yes	-0.3849	0.0181	NEG
Searches	3	LogitGAM	Yes	-0.3522	0.0186	NEG
Messages	1	LogitForest	Yes	-0.2270	0.0180	NEG
Messages	1	**RDITLogit**	Yes	-0.2656	0.0185	NEG
Messages	1	LogitLinear	Yes	-0.2977	0.0167	NEG
Messages	1	LogitGAM	Yes	-0.2911	0.0167	NEG
Messages	2	LogitForest	Yes	-0.3580	0.0200	NEG
Messages	2	LogitLinear	Yes	-0.4588	0.0191	NEG
Messages	2	LogitGAM	Yes	-0.4436	0.0193	NEG
Messages	3	LogitForest	Yes	-0.3710	0.0220	NEG
Messages	3	LogitLinear	Yes	-0.5012	0.0216	NEG
Messages	3	LogitGAM	Yes	-0.4774	0.0221	NEG

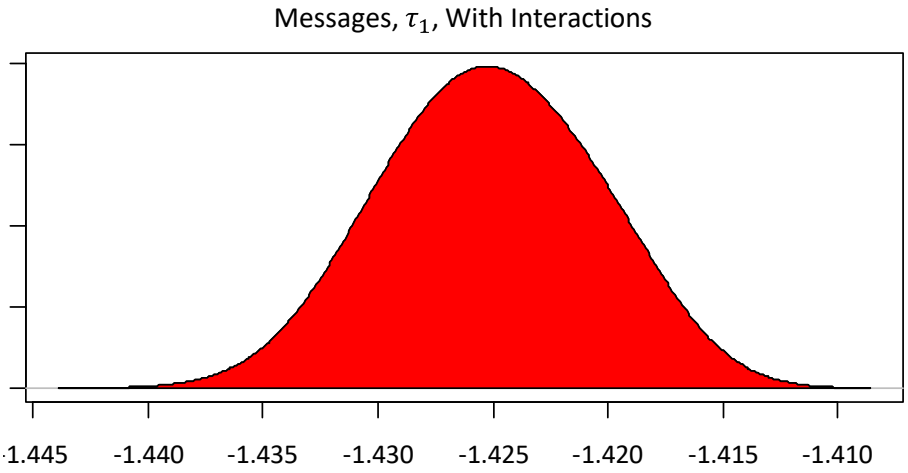
Appendix EC.10: Bootstrap Results for Treatment Effects, Linear Model (LM) with Interactions

DV	Week	Model	Bootstrap	Mean	SE	Result
Photos	1	LM	No	2.8045	0.0109	POS
Photos	1	LM	Yes	2.8051	0.0112	POS
Photos	2	LM	No	-0.1657	0.0006	NEG
Photos	2	LM	Yes	-0.1659	0.0012	NEG
Photos	3	LM	No	-0.2996	0.0005	NEG
Photos	3	LM	Yes	-0.2995	0.0010	NEG
Searches	1	LM	No	-0.2886	0.0050	NEG
Searches	1	LM	Yes	-0.2887	0.0060	NEG
Searches	2	LM	No	-3.3631	0.0117	NEG
Searches	2	LM	Yes	-3.3629	0.0122	NEG
Searches	3	LM	No	-3.3686	0.0110	NEG
Searches	3	LM	Yes	-3.3690	0.0116	NEG
Messages	1	LM	No	-1.4257	0.0059	NEG
Messages	1	LM	Yes	-1.4261	0.0070	NEG
Messages	2	LM	No	-2.4247	0.0121	NEG
Messages	2	LM	Yes	-2.4246	0.0143	NEG
Messages	3	LM	No	-1.9214	0.0047	NEG
Messages	3	LM	Yes	-1.9211	0.0057	NEG

Appendix EC.11: Bootstrap Distributions for Mean Treatment Effect, Linear Model (LM) with Interactions







Appendix EC.12: Comparative Covariate Effects First Post-Treatment Week

DV	Model	Coefficient	Interactions	Mean	SE	Result
Photos	LM	Age	Yes	0.1390	0.0068	POS
Photos	RLM	Age	Yes	0.1285	0.0067	POS
Photos	GAM	Age	—	—	—	—
Photos	RDITLinear	Age	Yes	0.0113	0.0042	POS
Photos	LogitForest	Age	Yes	0.1155	0.0063	POS
Photos	RDITLogit	Age	Yes	0.0037	0.0013	POS
Photos	LogitLinear	Age	Yes	-0.0020	0.0030	ns
Photos	LM	Cohort	Yes	-0.0037	0.0090	ns
Photos	RLM	Cohort	Yes	-0.0224	0.0088	NEG
Photos	GAM	Cohort	—	—	—	—
Photos	LogitForest	Cohort	Yes	-0.0370	0.0083	NEG
Photos	LogitLinear	Cohort	Yes	-0.0290	0.0050	NEG
Photos	LM	Married	Yes	0.5287	0.0238	POS
Photos	RLM	Married	Yes	0.4792	0.0232	POS
Photos	GAM	Married	Yes	0.5300	0.0238	POS
Photos	RDITLinear	Married	Yes	0.2073	0.0995	POS
Photos	LogitForest	Married	Yes	0.3581	0.0216	POS
Photos	RDITLogit	Married	Yes	0.4178	0.0371	POS
Photos	LogitLinear	Married	Yes	0.3280	0.0690	POS
Photos	LM	Private	Yes	-1.1537	0.0320	NEG
Photos	RLM	Private	Yes	-1.0142	0.0312	NEG
Photos	GAM	Private	Yes	-1.1466	0.0321	NEG
Photos	RDITLinear	Private	Yes	2.3240	0.0949	POS
Photos	LogitForest	Private	Yes	-0.8899	0.0311	NEG
Photos	RDITLogit	Private	Yes	0.4447	0.0747	POS
Photos	LogitLinear	Private	Yes	-0.1180	0.1810	ns
Searches	LM	Age	Yes	-0.0307	0.0031	NEG
Searches	RLM	Age	Yes	-0.0246	0.0030	NEG
Searches	GAM	Age	—	—	—	—
Searches	RDITLinear	Age	Yes	0.0861	0.0108	POS
Searches	LogitForest	Age	Yes	-0.0362	0.0064	NEG
Searches	RDITLogit	Age	Yes	0.0059	0.0006	POS
Searches	LogitLinear	Age	Yes	-0.0020	-0.0010	NEG
Searches	LM	Cohort	Yes	0.1148	0.0041	POS
Searches	RLM	Cohort	Yes	0.1261	0.0040	POS
Searches	GAM	Cohort	—	—	—	—
Searches	LogitForest	Cohort	Yes	0.3314	0.0086	POS
Searches	LogitLinear	Cohort	Yes	-0.0110	-0.0020	NEG

DV	Model	Coefficient	Interactions	Mean	SE	Result
Searches	LM	Married	Yes	-0.1033	0.0109	NEG
Searches	RLM	Married	Yes	-0.0853	0.0105	NEG
Searches	GAM	Married	Yes	-0.1000	0.0108	NEG
Searches	RDITLinear	Married	Yes	1.0421	0.2458	POS
Searches	LogitForest	Married	Yes	-0.1006	0.0222	NEG
Searches	RDITLogit	Married	Yes	0.1505	0.0142	POS
Searches	LogitLinear	Married	Yes	0.0410	0.0290	ns
Searches	LM	Private	Yes	0.1575	0.0147	POS
Searches	RLM	Private	Yes	0.1038	0.0142	POS
Searches	GAM	Private	Yes	0.1557	0.0145	POS
Searches	RDITLinear	Private	Yes	2.0169	0.2996	POS
Searches	LogitForest	Private	Yes	-0.0011	0.0297	ns
Searches	RDITLogit	Private	Yes	-0.0979	0.0202	NEG
Searches	LogitLinear	Private	Yes	-0.0590	0.0410	ns
Messages	LM	Age	Yes	-0.0387	0.0037	NEG
Messages	RLM	Age	Yes	-0.0301	0.0029	NEG
Messages	GAM	Age	—	—	—	—
Messages	RDITLinear	Age	Yes	0.0139	0.0093	ns
Messages	LogitForest	Age	Yes	-0.0622	0.0060	NEG
Messages	RDITLogit	Age	Yes	0.0092	0.0007	POS
Messages	LogitLinear	Age	Yes	0.0010	-0.0010	ns
Messages	LM	Cohort	Yes	-0.0987	0.0049	NEG
Messages	RLM	Cohort	Yes	-0.0229	0.0038	NEG
Messages	GAM	Cohort	—	—	—	—
Messages	LogitForest	Cohort	Yes	0.0713	0.0079	POS
Messages	LogitLinear	Cohort	Yes	0.0010	-0.0030	ns
Messages	LM	Married	Yes	-0.1603	0.0128	NEG
Messages	RLM	Married	Yes	-0.1113	0.0101	NEG
Messages	GAM	Married	Yes	-0.1638	0.0127	NEG
Messages	RDITLinear	Married	Yes	0.4394	0.2049	POS
Messages	LogitForest	Married	Yes	-0.1510	0.0209	NEG
Messages	RDITLogit	Married	Yes	0.2546	0.0180	POS
Messages	LogitLinear	Married	Yes	0.0690	0.0360	ns
Messages	LM	Private	Yes	0.2892	0.0173	POS
Messages	RLM	Private	Yes	0.1832	0.0136	POS
Messages	GAM	Private	Yes	0.2851	0.0171	POS
Messages	RDITLinear	Private	Yes	3.1362	0.2439	POS
Messages	LogitForest	Private	Yes	0.2507	0.0291	POS
Messages	RDITLogit	Private	Yes	-0.1845	0.0258	NEG
Messages	LogitLinear	Private	Yes	-0.1560	0.0540	NEG

Appendix EC.13: Granger Causality and Kolmogorov-Smirnov Test Results

Activity	Cohort	Res DF	F	p	KS DStat	KS p
Photos	3	21	1020.0	0	0.0870	0.999999
Photos	4	20	1345.9	0	0.0455	1
Photos	5	19	1452.6	0	0.0476	1
Photos	6	18	2123.3	0	0.0500	1
Photos	7	17	1752.6	0	0.0526	1
Photos	8	16	929.4	0	0.0556	1
Photos	9	15	1527.9	0	0.0588	1
Photos	10	14	3826.2	0	0.0625	1
Photos	11	13	1780.6	0	0.0667	1
Photos	12	12	2264.6	0	0.0714	1
Photos	13	11	2150.2	0	0.0769	1
Photos	14	10	942.7	0	0.0833	1
Photos	15	9	2585.8	0	0.0909	1
Photos	16	8	2480.3	0	0.1000	1
Photos	17	7	1009.2	0	0.1111	1
Photos	18	6	1504.7	0	0.1250	1
Photos	19	5	149.7	0.0003	0.1429	1
Photos	20	4	2070.2	0	0.1667	1
Photos	21	3	53.5	0.0182	0.2000	1
Searches	3	21	1717.4	0	0.0435	1
Searches	4	20	3489.8	0	0.0455	1
Searches	5	19	1248.6	0	0.0476	1
Searches	6	18	3359.0	0	0.0500	1
Searches	7	17	1858.5	0	0.0526	1
Searches	8	16	4790.3	0	0.0556	1
Searches	9	15	3168.8	0	0.0588	1
Searches	10	14	4804.6	0	0.0625	1
Searches	11	13	1468.1	0	0.0667	1
Searches	12	12	3171.0	0	0.0714	1
Searches	13	11	1668.9	0	0.0769	1
Searches	14	10	1510.9	0	0.0833	1
Searches	15	9	5698.0	0	0.0909	1
Searches	16	8	1983.3	0	0.1000	1
Searches	17	7	1119.4	0	0.1111	1
Searches	18	6	1585.4	0	0.1250	1
Searches	19	5	582.7	0	0.1429	1
Searches	20	4	359.8	0.0003	0.1667	1
Searches	21	3	117.6	0.0084	0.2000	1

Messages	3	21	2542.5	0	0.0435	1
Messages	4	20	1584.6	0	0.0455	1
Messages	5	19	1656.6	0	0.0476	1
Messages	6	18	4050.3	0	0.0500	1
Messages	7	17	2962.7	0	0.0526	1
Messages	8	16	5196.0	0	0.0556	1
Messages	9	15	2811.5	0	0.0588	1
Messages	10	14	6059.6	0	0.0625	1
Messages	11	13	6745.4	0	0.0667	1
Messages	12	12	2746.7	0	0.0714	1
Messages	13	11	3810.3	0	0.0769	1
Messages	14	10	3913.8	0	0.0833	1
Messages	15	9	10279.7	0	0.0909	1
Messages	16	8	3663.6	0	0.1000	1
Messages	17	7	10057.9	0	0.1111	1
Messages	18	6	2466.6	0	0.1250	1
Messages	19	5	744.1	0	0.1429	1
Messages	20	4	1585.6	0	0.1667	1
Messages	21	3	372.5	0.0027	0.2000	1