Harnessing Natural Experiments to Quantify the Causal Effect of Badges

Tomasz Kusmierczyk\(^1\) and Manuel Gomez-Rodriguez\(^2\)

\(^1\)Norwegian University of Science and Technology, tomaszku@idi.ntnu.no
\(^2\)Max Planck Institute for Software Systems, manuelgr@mpi-sws.org

Abstract

A wide variety of online platforms use digital badges to encourage users to take certain types of desirable actions. However, despite their growing popularity, their causal effect on users’ behavior is not well understood. This is partly due to the lack of counterfactual data and the myriad of complex factors that influence users’ behavior over time. As a consequence, their design and deployment lacks general principles.

In this paper, we focus on first-time badges, which are awarded after a user takes a particular type of action for the first time, and study their causal effect by harnessing the delayed introduction of several badges in a popular Q&A website. In doing so, we introduce a novel causal inference framework for badges whose main technical innovations are a robust survival-based hypothesis testing procedure, which controls for the utility heterogeneity across users, and a bootstrap difference-in-differences method, which controls for the random fluctuations in users’ behavior over time. We find that first-time badges steer users’ behavior if the utility a user obtains from taking the corresponding action is sufficiently low, otherwise, the badge does not have a significant effect. Moreover, for badges that successfully steered user behavior, we perform a counterfactual analysis and show that they significantly improved the functioning of the site at a community level.

1 Introduction

In recent years, social media sites and online communities have increasingly relied on digital badges to reward their users for different types of online behavior. Similarly as their physical counterpart, digital badges have been used both as a reputation mechanism, summarizing the skills and accomplishments of the users who receive them, and as an incentive mechanism, encouraging users to take certain type of desirable actions.

The promise of digital badges is that automated fine-grained monitoring and greater degree of control will help refine their design as incentive mechanisms, increasing users’ engagement and improving the functioning of the corresponding online platform. However, to fulfill this promise, it is necessary to better understand their causal effect on the online behavior of the users who may receive them—identify when and why they are (not) able to steer their behavior.

In this paper, we focus on first-time badges, which are awarded after a user takes a particular type of action for the first time\(^1\) and study their causal effect by harnessing several natural experiments in Stack Overflow\(^2\) a popular Q&A website. Despite their simplicity, we need to tackle several challenges, which require careful reasoning:

---

\(^1\)This work was done during Tomasz Kusmierczyk’s internship at the Max Planck Institute for Software Systems.

\(^2\)First-time badges are the simplest type of threshold badges, which are awarded after a user has taken an action a pre-specified number of times [15][26].

\(^2\)https://stackoverflow.com
— Measuring progress towards the badge: since first-time badges are awarded after performing just one single action, the action count does not provide a direct measure of progress towards the badge. This is in contrast with (non-binary) threshold badges, which were typically the focus of previous work [1, 13, 25].

— Utility heterogeneity: the utility each user obtains from taking an action differs wildly due to, e.g., user’s intrinsic motivation, the target of the action, or other users’ actions. As a consequence, the times users take to perform an action for the first time spans a large range of values.

— Random temporal changes: one can frequently observe random fluctuations in users’ behavior over time due to many different complex factors. To assess the strength of the causal effect induced by a badge, it is necessary to control for these random fluctuations. We address the above mentioned challenges by developing a novel causal inference framework for first-time badges, especially designed for our problem setting. Our framework avoids modeling the mechanisms underlying individual user actions and instead adopts a data-driven approach based on survival analysis and statistical hypothesis testing. At the heart of our approach there are two technical innovations: (i) a robust survival-based hypothesis testing procedure, inspired by the discrete choice literature on latent variable models [23, 6], which allows us to account for the utility heterogeneity; and, (ii) a bootstrap difference-in-differences method, inspired by the economics literature on natural experiments [16, 17, 21], which allows us to control for the random fluctuations in users’ behavior over time.

In contrast with recent empirical studies on threshold badges [4, 18, 19, 25], which assume or conclude that badges (always) steer users’ behavior, we do not find statistically significant causal evidence to back up this assumption (or conclusion) in all first-time badges. Instead, we provide strong empirical evidence of a more subtle picture. First-time badges steer users’ behavior if the utility a user obtains from taking the corresponding action is sufficiently low, otherwise, the badge does not have a significant effect. Moreover, we hypothesize that this may be also the case for other types of badges, e.g., non-binary threshold badges, and thus argue that the user utilities should be carefully considered on the design and deployment of badges. Finally, for badges that successfully steered user behavior, we go a step further and, using a survival-based counterfactual analysis, show that they significantly improve the functioning of the site at a community level.

**Related work.** Our work contributes to the growing literature on badges [2, 4, 7, 13, 25, 18, 19, 11], which can be broadly divided into theoretical and empirical studies.

Theoretical studies on badges [13, 25, 11] analyze the effect of badges on users’ behavior under stylized models of badges, which make strong assumptions, often without empirical support. Moreover, they typically ignore the (inherent) utility a user receives from taking the action the badge rewards—the action payoff and cost. In contrast, in our work, we avoid making strong assumptions about the mechanisms underlying individual user actions and instead adopt a data-driven approach, which enable us to account for the utility a user obtains from taking an action.

Empirical studies on badges [4, 18, 19] have mainly focused on threshold badges, where the action count provides a direct measure of progress towards the badge. In this context, several authors [18, 4] have provided empirical evidence in favor of the goal-gradient hypothesis, which posits users increase their engagement as they get closer to earning a badge. However, most of these studies did not have access to control groups, which would have allowed them to assess users’ behavior in the absence of a badge and control for random fluctuations in users’ behavior over time. Therefore, they are unable to identify when or why different types of threshold badges are able to steer users’ behavior. Two notable exceptions are by Abramovich et al. [2] and by Bornfeld et al. [7]. The former carried out a small controlled experiment in an educational setting and showed that the degree of success of a badge at steering a learner’s behavior depends on her ability and motivation. The latter has been concurrently conducted with our work and it also leverages natural experiments in the context of badges. However, in contrast to our work, they rely on standard statistical tests on aggregated counts, account for the temporal fluctuations in community functioning in an ad-hoc manner, and ignore the utility heterogeneity across users. As a consequence, they are unable to conclude whether badges had a (significant) causal effect at a user’s level and they do not shed light on when and why badges are (not) able to steer users’ behavior.

In recent years, natural experiments [17, 21], difference-in-difference designs [10, 16] and propensity score matching [9, 22] have been increasingly used to identify causal effects from observational data in online
settings, e.g., social influence \cite{5, 3, 8, 15} or network formation \cite{14, 20}. However, together with Bornfeld et al. \cite{7}, the present work is the first that leverage natural experiments to quantify causal effects in the context of badges.

2 Data description

Our Stack Overflow dataset comprises of all individual timestamped actions performed by all users from the site’s inception from July 31, 2008 to September 14, 2014, which allow us to track the complete sequence of actions users take.

**First-time badges: natural experiments.** There are a great variety of badges, which reward users for different types of behaviors. In this work, we focus on first-time badges, which are awarded after a user takes a particular type of action for the first time, and identify those that were introduced some time after the site’s inception. The delayed introduction of these badges can be thought of as natural experiments \cite{17, 21}. Figure 1 illustrates an example of such badge.

More specifically, we select three first-time badges that reward actions whose utilities to the users are clearly different:

— **Tag Editor badge:** Stack Overflow users can include tags on questions (or answers) to concisely describe their content. In July 2010, Stack Overflow enabled the creation of tag wikis by the community, which aim to provide a description of all used tags. Shortly afterwards, it introduced a badge called Tag Editor, awarded after a user edits a tag wiki for the first time, to encourage users to edit tag wikis. To ensure the quality of the wiki tags, only users with at least a reputation level of 1,500 could (initially) edit a tag wiki\footnote{In February 2011, Stack Overflow lowered the minimum reputation level to 100 and thus the characteristics of the population that could earn the badge changed. Therefore, in our analysis, we only consider data up to January 2010.}. Finally, note that a user obtains a low utility from editing a wiki tag—it requires some effort and she only receives the intangible reward of helping the community. Moreover, the more uncommon a tag is, the least this intangible reward may be.

— **Promoter badge:** When a Stack Overflow user does not receive a satisfactory answer to one of her questions, she can offer a bounty to reward, in the form of reputation points, the user who would provide such a satisfactory answer\footnote{A user can also offer a bounty to a user after she has provided an answer, as a thank you gift, however, that usage is rarer.}. In July 2010, Stack Overflow introduced a badge called Promoter, awarded after a user offers a bounty for an answer to one of her questions for the first time, to encourage users to offer more bounties. Only users with at least a reputation of 75 points can offer a bounty. In contrast with editing a wiki tag, a user obtains a high utility from offering a bounty—it requires little effort and she may receive an answer to a question she is personally interested in, however, it entails a cost in terms of the reputation she transfers to the user providing the answer.

---

Figure 1: Time when users first edited a tag wiki (user action time, \( t \)) against time when they became eligible to edit tag wikis (user start time, \( s \)). The horizontal black line denotes the time when the Tag editor badge was introduced, which is awarded after a user edits a tag wiki for the first time.
— **Investor badge**: Stack Overflow users can also offer bounties to receive a satisfactory answer to a question that has been asked by another user. In July 2010, Stack Overflow introduced a first-time badge called Investor to encourage users to offer more bounties for answers to other users’ questions. Similarly as in the Promoter badge, only users with at least a reputation of 75 points can offer a bounty for an answer to a question asked by another user. However, in this case, a user may obtain a lower utility from offering a bounty for an answer to other user’s question than her own—on the one hand, she may less interested in an answer since she did not originally ask the question and, on the other hand, the question may have already a (relatively) satisfactory answer when she found it.

3 Testing the effectiveness of badges

In this section, we first formalize the problem setting. Then, we introduce two survival-based hypothesis testing procedures of increasing statistical power, which we use in a novel bootstrap difference-in-differences method, especially designed for our problem setting. Finally, we evaluate the effectiveness of the overall framework using a variety of synthetic experiments.

**Problem setting.** Given an action of interest \( a \), we record the behavior of each user during an observation window \([0, T]\) as a tuple

\[
e := (s_u, t_u, v_u),
\]

which means that user \( u \) becomes eligible to perform the action at time \( s_u \), she performs the action at time \( t_u \), and obtains a utility \( v_u \), which is often intangible. If a user does not perform the action during the observation window \([0, T]\), we set the action time to \( t_u = \infty \), however, this does not imply she will never perform the action. Moreover, we assume a first-time badge \( b \) is introduced at time \( \tau \in [0, T] \) to incentivize users to take action \( a \). That means, after time \( \tau \), a user receives badge \( b \) the first time she takes action \( a \), potentially increasing its corresponding utility \( v_u \).

Given the above setup, our goal is then to assess to which extent the introduction of the badge changes users’ behavior, as measured by the time users take to perform the action for the first time, i.e., \( t_u - s_u \). Next, we introduce two survival-based hypothesis testing procedures of increasing statistical power and then describe our bootstrap difference-in-differences method.

**Basic survival-based hypothesis testing.** Given an action of interest \( a \), we model the time \( t_u \) when a user \( u \) takes action \( a \) using a survival process \( \Xi \). Following the literature on temporal point processes, we represent such survival process as a binary counting process \( \mathcal{N}_u(t) \in \{0,1\} \), which becomes one when the user performs the action for the first time. Then, we characterize this counting process using its corresponding intensity \( \lambda_u(t) \), i.e., \( \mathbb{E}[d\mathcal{N}_u(t)] = \lambda_u(t)dt \), which we define as follows:

\[
\lambda_u(t) = \begin{cases} 
0 & \text{if } t < s_u \\
\lambda_0 & \text{if } s_u \leq t < \tau \\
\lambda_1 & \text{otherwise}
\end{cases}
\]

where \( \lambda_0 \) and \( \lambda_1 \) are parameters shared across all users, which depend on the (intangible) utility users obtain from talking the action.

Under this model, the null hypothesis \( H_0 \), i.e., the badge did not have an effect, corresponds to \( \lambda_0 = \lambda_1 \geq 0 \) and the alternative hypothesis \( H_1 \) is \( \lambda_0 \neq \lambda_1 \) with \( \lambda_0 \geq 0 \) and \( \lambda_1 \geq 0 \). Moreover, given the behavior of \( n \) users, the maximum likelihood estimators of the model parameters, \( \hat{\lambda}_0 \) and \( \hat{\lambda}_1 \), can be computed analytically. In particular, under the null hypothesis, they are readily given by:

\[
\hat{\lambda}_0 = \hat{\lambda}_1 = \frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq T)}{\sum_{u \in [n]} (\min(t_u, T) - s_u)},
\]
and under the alternative hypothesis, they are given by:

\[
\hat{\lambda}_0 = \frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq \tau) - \hat{s}_u \mathbb{I}(s_u < \tau)}{\sum_{u \in [n]} (\min(t_u, \tau) - s_u)(s_u < \tau)}
\]

\[
\hat{\lambda}_1 = \frac{\sum_{u \in [n]} \mathbb{I}(\tau_u < t_u \leq T)}{\sum_{u \in [n]} \mathbb{I}(\tau_u < t_u \leq T - \tau)(t_u > \tau)}
\]

where \(\mathbb{I}(\cdot)\) is the indicator function and all the sums are over eligible users. Then, we can use a standard log-likelihood ratio (LLR) as test statistic \([12]\). Moreover, since the null model is nested in the alternative model, \(i.e., \theta_0 \in \{\theta_0, \theta_1\}\), using Wilks’ theorem \([24]\), it asymptotically holds that \(2LLR \sim \chi^2_1\) under the null hypothesis. Thus, we can easily find an approximate \(p\)-value, \(i.e., p \approx 1 - \chi^2_1(2LLR)\).

**Robust survival-based hypothesis testing.** The survival-based hypothesis testing procedure described in the previous section assumes the model parameters are shared across all users and, by doing so, it ignores the utility heterogeneity across users. Inspired by the discrete choice literature on latent variable models \([23, 6]\), we account for the utility heterogeneity by considering different latent parameters per user, but sampled from the same distributions, \(i.e.,\)

\[
\lambda_u(t) = \begin{cases} 
0 & \text{if } t < s_u \\
\lambda_0(u) & \text{if } s_u \leq t < \tau \\
\lambda_1(u) & \text{otherwise}
\end{cases}
\]

\(\lambda_0(u) \sim \text{Gamma}(k_0, r)\)

\(\lambda_1(u) \sim \text{Gamma}(k_1, r)\) \quad (3)

where \(k_0, k_1\) are shape parameters and \(r\) is a rate parameter. Here, note that \(E[\lambda_0] = k_0/r\) and \(E[\lambda_1] = k_1/r\).

Then, we define the null and alternative hypothesis in terms of the shape parameters, \(i.e.,\)

\[
H_0 : k_0 = k_1 \geq 0
\]

\[
H_1 : k_0 \neq k_1, k_0 \geq 0, k_1 \geq 0
\]

Moreover, given the behavior of \(n\) users, we can estimate the shape parameters using maximum likelihood estimation, integrating out the latent parameters \(\lambda_0(u)\) and \(\lambda_1(u)\), and estimate the rate parameter by cross validation. More specifically, under the null hypothesis, the shape parameters are given by:

\[
\hat{k}_0 = \hat{k}_1 = -\frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq T)}{\sum_{u \in [n]} \log \left( \frac{r}{\tau + \min(t_u, \tau) - s_u} \right)}
\]

and under the alternative hypothesis, they are given by:

\[
\hat{k}_0 = -\frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq \tau)}{\sum_{u \in [n]} \log \left( \frac{r}{\tau + \min(t_u, \tau) - s_u} \mathbb{I}(s_u \leq \tau) \right)}
\]

\[
\hat{k}_1 = -\frac{\sum_{u \in [n]} \mathbb{I}(\tau_u < t_u \leq T)}{\sum_{u \in [n]} \log \left( \frac{r}{\tau + \min(t_u, \tau) - \tau} \mathbb{I}(t_u > \tau) \right)}
\]

Similarly as in the basic survival model, we can then use a standard log-likelihood ratio (LLR) as test statistic. However, in this case, the null model is not nested in the alternative model and, as a consequence, we cannot use Wilks’ theorem to find a \(p\)-value. Instead, we will rely on the following bootstrap difference-in-difference method to find a robust empirical estimate of the distribution of the test statistic under the null hypothesis.
Figure 2: Our bootstrap difference-in-difference method. The treatment population (left) consists of users whose start time \( s_u \) lies in a window of size \( w \) around the time \( \tau \) the badge is introduced. Each control population \( i \) (right) consists of users whose start time \( s_u \) lies in a window of size \( w \) around the time \( \tau_i \) a virtual badge is introduced. The method runs hypothesis testing on both the treatment population and all control populations and then compares the test statistic (e.g., p-value) of the treatment population with the empirical distribution of the test statistic of the control populations.

**Bootstrap difference-in-differences method.** Given an action of interest \( a \), its corresponding first-time badge \( b \) with introduction time \( \tau \), the behavior of \( n \) users with respect to \( a \), i.e., \( D_a = \{(s_u, t_u, v_u)\}_{u \in [n]} \), and a model-based hypothesis testing procedure, we design the following bootstrap difference-in-difference method to find a robust empirical estimate of the distribution of the corresponding test statistic under the null hypothesis, which accounts for the temporal fluctuations in users’ behavior:

I. We select all users whose start time \( s_u \in [\tau - w/2, \tau + w/2] \), where \( w \geq 0 \) is a given parameter. Then, we run the model fitting procedure of choice on that subset of users, the treatment population, and obtain a test statistic, e.g., \( LLR_\tau \).

II. We introduce a set of virtual badges \( \mathcal{V} \) at a times \( \tau_i \in [w/2, \tau - w] \cup [\tau + w, T - w/2] \), picked uniformly at random (in practice, one can use a sliding window), where \( w \geq 0 \) is the same given parameter as in the first step. Then, for each virtual badge \( i \in \mathcal{V} \), we select users whose start time \( s_u \in [\tau_i - w/2, \tau_i + w/2] \).

Finally, we run the model fitting as in the previous step on each of these subsets of users, the control populations, and obtain a test statistic value per virtual badge, e.g., \( LLR_{\tau_i} \). As a result, we can estimate an empirical cumulative density function (cdf) of the test statistic under the null hypothesis, \( F_{LLR}(LLR) \), which is robust to temporal fluctuations in users’ behavior.

III. We measure the strength of the change induced by the badge by means of the probability that the test statistic of the control populations (for which the null hypothesis holds by design) is larger than the test statistic of the treatment population, \( p := F_{LLR}(LLR_T) \).

The above bootstrap difference-in-differences method, which we also illustrate in Figure 2, equips us with a robust empirical estimate of the distribution of the test statistic under the null hypothesis \( F_{LLR}(LLR) \) and a p-value \( p = F_{LLR}(LLR_T) \), which accounts for the temporal fluctuations in users’ behavior and allows us to reject the null hypothesis with higher confidence. The main assumption needed for the above method to be valid (e.g., the empirical estimate of the test statistic distribution under the null hypothesis to be accurate) is that the treatment and control populations have similar characteristics. In other words, the process governing the exposures should resemble random assignment.

**Framework evaluation.** In this section, we compare the effectiveness of the basic survival model with the theoretical distribution (\( \chi^2_2 \)) of the LLR under the null hypothesis (“basic theoretical”) and the basic and robust survival models with the empirical distribution (\( F_{LLR} \)) of the LLR under the null hypothesis, as estimated by the proposed difference-in-differences bootstrap method (“basic bootstrap” and “robust bootstrap”, respectively). More specifically, we proceed as follows.

First, we simulate the behavior of \( n = 10,000 \) users during a time interval \([0, T] \), where \( T = 360 \). For each user, we draw her starting times \( s_u \) uniformly at random, \( s \sim U[0, T] \), and her action time \( t \) from an intensity \( \lambda_u(t)(1 + at) \), where \( \lambda_u(t) \) is given by Eq. 3 and \( a = 0.001 \). Moreover, in Eq. 3 we set the badge
introduction time to \( \tau = T/2 \), the rate parameter to \( r = 10 \), and consider different badge strength values, \( i.e., E(\lambda_1(u))/E(\lambda_0(u)) = k_1/k_0 \in \{1.0, 1.25, \ldots, 10, 100\} \), where \( k_1/k_0 = 1.0 \) is equivalent to not introducing a badge. Note that the term \((1 + \alpha t)\) imposes a global linear trend, which is often observed in real data\(^5\). For each configuration, we run 100 independent simulations.

Then, we run the above methods (“basic theoretical”, “basic bootstrap”, and “robust bootstrap”) on data from each of the independent simulations and measure their effectiveness in terms of two metrics: average \(p\)-value and rejection probability of the null hypothesis \( H_0 \) at \( p = 0.05 \). Figure 3 summarizes the results, which show that the robust bootstrap has a superior performance: it is more likely to reject \( H_0 \) when a badge is introduced (\(i.e., k_1/k_0 > 1.0\)) while it is equally likely not to reject \( H_0 \) when a badge is not introduced (\(i.e., k_1/k_0 = 1.0\)).

![Figure 3: Performance of our causal inference framework on synthetic data. The left panel shows the average \(p\)-value and rejection probability of the null hypothesis \( H_0 \) against effect strength \( k_1/k_0 \), where lower (higher) is better for \( k_1/k_0 > 1 \) (\(k_1/k_0 = 1\)). The right panel shows the rejection probability of the null hypothesis \( H_0 \) at \( p = 0.05 \) against effect strength \( k_1/k_0 \), where higher (lower) is better for \( k_1/k_0 > 1 \) (\(k_1/k_0 = 1\)).](image)

### 4 Do badges work?

In this section, we apply our causal inference framework to the three first-time badges described in Section 2. Figure 4 summarizes the results by means of:

(i) Test statistic over time for the basic and robust survival models, \( i.e., LLR_\tau \) and \( LLR_\tau \) against \( \tau_1 \) and \( \tau \).

(ii) Empirical distribution of the test statistic under \( H_0 \) and \( p\)-value for the robust survival model, \( i.e., F_{LLR}(LLR) \) and \( F_{LLR}(LLR_\tau) \).

(iii) Average intensities for first-time action under robust survival model, \( i.e., \hat{k}_0 \) before \( \tau \) and \( \hat{k}_1 \) after \( \tau \), using a sliding window of length \( w = 60 \) days.

Overall, the results suggest that the Tag editor and Investor badges were successful—they had a significant causal effect on users’ behavior (\( p = 0.004 \) and \( p = 0.017 \), respectively). In contrast, the Promoter badge was unsuccessful—it did not have a significant causal effect (\( p = 0.309 \)). Moreover, a detailed analysis also reveals several interesting insights, which we will further expand in Sections 5 and 6.

First, the actions rewarded by the two successful badges were rare by the time the badges got introduced. For example, in the case of the Tag editor badge, only 100 tag wiki edits had been performed, however, there were \( \sim 6,500 \) users who were eligible to perform edits. In the case of the Investor badge, only 40 bounties had been offered for an answer to other users’ questions. In contrast, the action rewarded by the Promoter badge—offering a bounty for an answer to the users’ own questions—was much more common by the time the badge got introduced. As a consequence, the average intensity for the Promoter badge was an order of magnitude higher than the intensities corresponding to the Tag editor or the Investor badge.

\(^5\)We obtain quantitatively similar results in the absence of a linear trend, \( i.e., \alpha = 0 \).
Second, the introduction of the Tag editor and Investor badges was followed by an increase on the average intensity of the corresponding first-time action of more than 4·, from $k_0 \approx 2 \cdot 10^{-4}$ to $k_1 \approx 8 \cdot 10^{-4}$ for Tag editor badge and from $k_0 \leq 5 \cdot 10^{-5}$ to $k_1 \approx 2 \cdot 10^{-4}$ for Investor badge. Equivalently, the average time a user takes to perform the actions for the first time was reduced by 75%. Moreover, this change in user behavior did not vanish over time, as shown in the rightmost column.

Finally, in the case of the Investor badge, we find a transient increase on the average intensity of bounties to other users’ questions around October-November 2010, which is statistically significant. Upon investigation, we notice that several users discovered ways of benefiting from offering bounties around that time, triggering subsequent first-time uses of bounties by other users. Such discussions led to an increase on the minimum reputation one can transfer when offering a bounty.

5 Badges and utilities

In this section, we investigate the reasons why the Tag editor and Investor badges were successful at steering users’ behavior while the Promoter badge was unsuccessful. To this aim, we resort to game-theoretic
concepts such as user utilities, action payoffs and reservation values, and identify measurable proxies of some of these concepts for each of the above mentioned badges and actions.

**User utilities.** In the game theory literature \[11, 13, 24\], the utility a user obtains from performing an action is defined as the difference between the action payoff \( p \) and the cost of effort \( c \), i.e., \( v = p - c \). Moreover, the fact that participation is a voluntary, strategic choice—users have a choice about whether or not to perform an action—is often modeled via a reservation value \( \omega \) that the utility \( v \) must exceed in order for the user to perform the action. More specifically, if \( p - c < \omega \), the user will decide not to perform the action and, otherwise, she will perform it. In this context, a badge \( b \) is assumed to increase the utility a user obtains from performing the action, i.e., \( v = p - c + v_b \), where \( v_b \) is the badge value. Then, depending on the actual values of \( p, c, v_b \) and \( \omega \), one can argue that a badge will induce users to perform an action that, in the absence of a badge, would not perform.

However, in social media sites and online communities, the action payoffs, cost of effort, badge value and reservation values are typically intangible, hidden or ambiguously defined. As a consequence, our causal inference framework did not explicitly adopted the above model and instead used a data-driven approach based on survival analysis of items specific for particular badge, using only the observable temporal traces. In this section, however, we turn our attention towards the above stylized model, identify measurable proxies of the model parameters for each of the above mentioned badges and actions, and use them to investigate the reasons for the success or failure of badges at steering users’ behavior, as concluded by our framework.

**Proxies to user utilities.** We consider the following observable proxies for the utilities users obtain from editing a wiki tag and offering a bounty, respectively:

(a) **Tag popularity:** the more popular a tag is, the greater the utility \( v \) a user may obtain from editing its wiki tag—the user needs to put less effort to create a wiki on a popular tag and she receives the satisfaction of helping a larger part of the community.

(b) **Number of answers:** the higher (lower) the number of answers a question receives after (before) offering a bounty, the greater the utility \( v \) a user obtains from offering the bounty. Moreover, users offering a bounty for an answer to other user’s question may obtain less utility from the answers since they did not originally ask the question.

Given the above proxies, we proceeds as follows. In terms of tag wikis, we group tags by popularity (i.e., number of questions a tag was used on) and model the time the users take to create a wiki for a tag of a given popularity \( p \) as a survival process. Moreover, we characterize this process using an intensity \( \lambda_p(t) \), which we define as follows:

\[
\lambda_p(t) = \begin{cases} 
0 & \text{if } t < s \\
\lambda_0(p) & \text{if } s \leq t < \tau \\
\lambda_1(p) & \text{otherwise}
\end{cases}
\]  

Figure 5: Causal effect of the **Tag Editor** badge for tags with different utility value, as estimated by their popularity level.
where \( \lambda_0(p) \) and \( \lambda_0(p) \) are parameters shared across all tags with popularity \( p \), \( s \) is the time when the tag is first used in a question, and \( \tau \) is the time when the Tag Editor badge is introduced. Then, by comparing the maximum likelihood estimators of the model parameters, \( \lambda_0(p) \) and \( \lambda_1(p) \), for different popularity levels \( p \), we can assess the causal effect of the Tag Editor badge on tag wikis with different utility values.

In terms of bounties, we first group questions by the number of answers they received in the first two days since they were asked and then compare the additional number of answers they received after those first two days if a bounty was (not) offered in the second day. Moreover, for questions that received a bounty, we estimate the distribution of the number of answers they received before the bounty was offered both before and after the badges Promoter and Investor were introduced. By controlling for the number of answers, we can assess the causal effect of both badges for bounties with different utility values.

Results. Figure 5 summarizes the results for the Tag Editor badge, which shows the higher the popularity (utility) of a tag, the weaker the causal effect of the badge introduction. In other words, the introduction of the badge steered users to create tag wikis for less popular, low utility tags.

Figure 6 summarizes the results for the Promoter and Investor badges, which let us better understand their failure and success, respectively: (i) the number of answers a bounty triggers (i.e., its utility) increases with the number of answers the question has received in its absence (left panel); (ii) the introduction of the Promoter badge did not significantly change the users’ willingness to offer bounties to their own questions (right panel, top figure), in contrast, the Investor badge did change it for bounties offered to other users’ questions (right panel, bottom figure). This change was especially pointed for questions with zero answers before offering the bounty, in which the utility of offering a bounty is the lowest (left panel).

6 Do badges improve the community functioning?

In this section, we use a survival-based counterfactual analysis, in which we investigate what would have happened if the Tag Editor and Investor badges had not been introduced, to assess to which extent the badges improved the site functioning at a community level.

Counterfactual analysis. For the Tag editor, we assess the site functioning at a community level in terms of the number of new tag wikis over time and, for the Investor badges, we use the time to bounty and first answer across questions. More specifically, we proceed as follows.

— New tag wikis: we simulate the time the users take to create a new wiki for a tag of a given popularity \( p \) in the counterfactual world where the Tag editor badge is never introduced using the intensity defined by Eq. 4 with \( \lambda_0(p) = \lambda_1(p) = \hat{\lambda}_0(p) \), where \( \hat{\lambda}_0(p) \) is the maximum likelihood estimate for \( \lambda_0(p) \) in the true world.
The Tag Editor badge introduction improved the site functioning at a community level in terms of the number of new tag wikis (left panel; the number of new tag wikis increased) and the average popularity (rank) of the associated (right panel; new tags wikis were created for less popular tags). However, shortly after its introduction, the rate of creation of new tag wikis decreased.

---

A bounty can only be offered two days after the question has been asked. 

---

Figure 7: Tag wikis with and without the Tag Editor badge. Simulations means and 95%-CI shown.

Figure 8: Time to bounty and first answer with and without the Promoter and Investor badges.

Then, we compare the number of new tag wikis over time as well as the popularity of their associated wikis in the true world and in the simulated counterfactual world.

— Time to bounty and first answer: for questions that received a bounty, we model the time that the users take to offer the bounty as a survival process with associated intensity \( \lambda_b(t) \), which we define as follows:

\[
\lambda_b(t) = \begin{cases} 
0 & \text{if } t < s \\
\lambda_0(b) & \text{if } s \leq t < \tau \\
\lambda_1(b) & \text{otherwise}
\end{cases}
\]

where the parameter \( b \in \{0, 1\} \) denotes whether the badge is offered by the user asking the question or by another user, \( \{\lambda_i(b)\}_{i, b \in \{0, 1\}} \) are (four) parameters shared across all questions, \( s \) is the time since two days after the question is asked, and \( \tau \) is the time when the Promoter and Investor badges are introduced. For all questions, we model the time that the users take to provide the first answer also as a survival process with an associated intensity defined similarly as in Eq. 5, however, in this case, the parameter \( b \in \{-1, 0, 1\} \) denotes whether the question received a badge and, if so, whether it was offered by the user asking the question or by another user, and thus the model has six parameters. Then, we compare the maximum likelihood estimators of the above parameters to assess what would have happened if the Investor badge had not been introduced.

**Results.** Figure 7 summarizes the results for the Tag Editor badge. The badge introduction improved the site functioning at a community level in terms of the number of new tag wikis (left panel; the number of new tag wikis increased) and the average popularity (rank) of the associated (right panel; new tags wikis were created for less popular tags). However, shortly after its introduction, the rate of creation of new tag wikis decreased.
decreased to its original value and matched the rate of creation in the simulated counterfactual world. A plausible hypothesis is that the badge lifted the utility value of badges over the reservation value for tags of certain popularity. As a consequence, after these tags had a tag wiki, the effect of the badge diminished.

Figure 8 summarizes the results for the Investor badge, which show that the time to bounty and first answer for questions in which a bounty was offered by a user different than the user asking the question decreased (i.e., the intensities increased) after the Investor was introduced, improving the site functioning. In contrast, the time to first answer (and time to bounty) in questions without bounty (and with a bounty offered by the user asking the question) increased, suggesting that, in the counterfactual world where the Investor badge had not been introduced, the site functioning would have actually worsened.

7 Conclusions

Social media sites and online communities are dynamic environments where users change their behavior on a daily basis due to many complex factors. As a consequence, assessing the effectiveness of incentive mechanisms, which are ubiquitous among them, is challenging. In this work, we have focused on one of the simplest incentive mechanisms—first-time badges—and studied their effectiveness by developing a novel survival-based causal modeling framework, specially designed to harness the delayed introduction of several badges in a popular Q&A website.

Our work also opens up many interesting venues for future work. For example, it would be very useful to extend our framework to other types of badges, e.g., threshold badges. Badges are typically awarded to all users whose contributions exceed some predefined values, however, it would be interesting to also consider incentive mechanisms where users compete for a limited reward. Finally, in this work, we have focused on assessing the causal effect of (first-time) badges. A natural follow-up would be developing principled, effective methods to optimize their design.

Acknowledgements

We thank Utkarsh Upadhyay and Isabel Valera for useful discussions.

References

[1] O. Aalen, O. Borgan, and H. Gjessing. Survival and event history analysis: a process point of view. Springer Science & Business Media, 2008.

[2] S. Abramovich, C. Schunn, and R. M. Higashi. Are badges useful in education?: It depends upon the type of badge and expertise of learner. Educational Technology Research and Development, 61(2):217–232, 2013.

[3] T. Althoff, P. Jindal, and J. Leskovec. Online actions with offline impact: How online social networks influence online and offline user behavior. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.

[4] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. In Proceedings of the 22nd international conference on World Wide Web, 2013.

[5] S. Aral and D. Walker. Identifying influential and susceptible members of social networks. Science, 337(6092):337–341, 2012.

[6] M. Ben-Akiva, J. Walker, A. T. Bernardino, D. A. Gopinath, T. Morikawa, and A. Polydoropoulou. Integration of choice and latent variable models. Perpetual motion: Travel behaviour research opportunities and application challenges, pages 431–470, 2002.
[7] B. Bornfeld and S. Rafaeli. Gamifying with badges: A big data natural experiment on stack exchange. First Monday, 22(6), 2017.

[8] Y. Chen, Q. Wang, and J. Xie. Online social interactions: A natural experiment on word of mouth versus observational learning. Journal of marketing research, 48(2):238–254, 2011.

[9] R. H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics, 84(1):151–161, 2002.

[10] S. G. Donald and K. Lang. Inference with difference-in-differences and other panel data. The review of Economics and Statistics, 89(2):221–233, 2007.

[11] D. Easley and A. Ghosh. Incentives, gamification, and game theory: an economic approach to badge design. ACM Transactions on Economics and Computation, 4(3):16, 2016.

[12] R. V. Hogg and A. T. Craig. Introduction to mathematical statistics. Prentice Hall, 1995.

[13] N. Immorlica, G. Stoddard, and V. Syrgkanis. Social status and badge design. In Proceedings of the 24th international conference on World Wide Web, 2015.

[14] A. Z. Jacobs, S. F. Way, J. Ugander, and A. Clauset. Assembling thefacebook: Using heterogeneity to understand online social network assembly. In Proceedings of the ACM Web Science Conference, page 18. ACM, 2015.

[15] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24):8788–8790, 2014.

[16] M. Lechner. The estimation of causal effects by difference-in-difference methods. Foundations and Trends® in Econometrics, 4(3):165–224, 2011.

[17] B. D. Meyer. Natural and quasi-experiments in economics. Journal of business & economic statistics, 13(2):151–161, 1995.

[18] T. Mutter and D. Kundisch. Behavioral mechanisms prompted by badges: The goal-gradient hypothesis. In Proceedings of the 35th International Conference on Information Systems, 2014.

[19] H. Oktay, B. J. Taylor, and D. D. Jensen. Causal discovery in social media using quasi-experimental designs. In Proceedings of the First Workshop on Social Media Analytics, pages 1–9, 2010.

[20] T. Q. Phan and E. M. Airoldi. A natural experiment of social network formation and dynamics. Proceedings of the National Academy of Sciences, 112(21):6595–6600, 2015.

[21] M. R. Rosenzweig and K. I. Wolpin. Natural "natural experiments" in economics. Journal of Economic Literature, 38(4):827–874, 2000.

[22] E. A. Stuart. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1):1, 2010.

[23] J. L. Walker. Extended discrete choice models: integrated framework, flexible error structures, and latent variables. PhD thesis, Massachusetts Institute of Technology, 2001.

[24] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics, 9(1):60–62, 1938.

[25] J. Zhang, X. Kong, and P. S. Yu. Badge System Analysis and Design. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2016.