On the Causal Effect of Badges

Tomasz Kuśmierczyk
Norwegian University of Science and Technology
tomaszku@ntnu.no

Manuel Gomez-Rodriguez
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 first-time badges whose main technical innovations are a robust survival-based hypothesis testing procedure, which controls for the heterogeneity in the benefit users obtain from taking an action, and a bootstrap difference-in-differences method, which controls for the random fluctuations in users' behavior over time. Our results suggest that first-time badges steer users' behavior if the initial benefit a user obtains from taking the corresponding action is sufficiently low, otherwise, we do not find significant effects. 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. First-time badges are the simplest type of the highly popular threshold badges [4, 28, 35], which are awarded after a user has taken an action a pre-specified number of times. More specifically, we study the causal effect induced by first-time badges by harnessing several *natural experiments* in Stack Overflow[1], a popular Q&A website, consisting of the *delayed* introduction of a badge some time after the site's inception. Despite their simplicity, we need to tackle several challenges, which require careful reasoning:

— *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 [4, 28, 35].

— *Utility heterogeneity:* the benefit (or *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. As a consequence, 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, 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 [23, 25, 31], which allows us to control for the random fluctuations in users' behavior over time. Moreover, while our framework focuses on first-time badges, we argue that our methodological innovations will also shed light on more complex badges, *e.g.*, non-binary threshold badges.

In contrast with recent empirical studies on threshold badges [4, 28, 29, 35], which typically assume or conclude that, to some degree, badges always steer users' behavior, we do not find statistically significant evidence to back up this conclusion in all the first-time badges we considered. Instead, we provide empirical evidence of a more subtle picture. In particular, our results suggest that first-time badges steer users' behavior if the utility a user obtains from taking the corresponding action

---

[1]https://stackoverflow.com

is sufficiently low, otherwise, the badge does not seem to have a significant effect. Moreover, we hypothesize that this may be also the case for 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, 5, 7, 12, 19, 28, 29, 35], which can be broadly divided into theoretical and empirical studies.

Theoretical studies on badges [12, 19, 35] 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 [2, 4, 5, 7, 28, 29] have mainly focused on threshold badges, where the action count provides a direct measure of progress towards the badge. In this context, several authors [4, 28] 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, while other authors [2, 5] found that degree of success of a badge may depend on different complex factors. 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. A notable exception is by Bornfeld et al. [7], which has been concurrently and independently 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 users' behavior in an ad-hoc manner, and ignore the utility heterogeneity across users.

Moreover, badges can be viewed as a *gamification* mechanism, *i.e.*, a mechanism in which game-design elements are used in a non-gaming environment to encourage users to perform certain tasks [10, 32]. Therefore, our research adds up to the existing studies on practical implications of gamification for, *e.g.*, increasing user engagement in learning [15, 16] and e-learning [13, 27], shaping healthy behavioral patterns [14, 21, 24], systems design [17] or crowdsourcing [26].

Finally, natural experiments [25, 31], difference-in-difference designs [11, 23] and propensity score matching [9, 33] have been increasingly used to identify causal effects from observational data in online settings, *e.g.*, social influence [3, 6, 8, 22] or network formation [20, 30]. However, together with Bornfeld et al. [7], the present work is one of the first that leverage natural experiments to quantify causal effects in the context of badges.
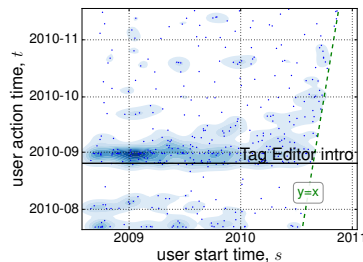
Figure 1: Time when users first edited a tag wiki (user action time, t) against time when they became eligible to edit tag wikis (user start time, s). The horizontal black line denotes the time when the `Tag editor` badge was introduced, which is awarded after a user edits a tag wiki for the first time.

2 DATA DESCRIPTION

Our Stack Overflow dataset[2] 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* [25, 31], where the treatment of earning a badge is assigned *as if random* among users. 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[3]. 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 since she will help a smaller part of the community.

—`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[4]. In July 2010, Stack Overflow introduced a badge called `Promoter`,

---

[2]Publicly available at https://archive.org/details/stackexchange

[3]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.

[4]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.

Table 1: General statistics on the number of users and actions in our dataset. The time when a badge is introduced is denoted by τ, the time when a user becomes eligible to perform an action is denoted by s_u and the time when a user performs the action is denoted by t_u.

	Tag Editor*	Investor	Promoter
Required reputation	1,500	75	75
Eligible users, $s_u < \tau$	6,396	46,148	46,148
Actions, $t_u < \tau$	93	4,984	205
Eligible users, $s_u \geq \tau$	2,095	297,481	297,481
Actions, $t_u \geq \tau$	471	30,396	7,830

*For `Tag Editor` badge we present statistic only until 2011-02-09 when the required reputation changed.

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.

— `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 another user's question than her own—on the one hand, she may be 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.

Table 1 provides general statistics on the number of eligible users and user actions in our dataset.

3 TESTING THE EFFECTIVENESS OF FIRST-TIME BADGES

In this section, we first formalize the problem setting, which includes at its core a natural experiment. Then, we introduce our causal inference framework for first-time badges, discussing the assumptions it requires and elaborating on its individual components. Finally, we evaluate its effectiveness using a variety of synthetic experiments and conclude with a discussion of its limitations.

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

$$e := (\overset{\text{start time}}{\underset{\downarrow}{s_u}}, \underset{\underset{\uparrow}{\text{action time}}}{t_u}, \overset{\text{utility}}{\underset{\downarrow}{v_u}}), \qquad (1)$$

which means that user u becomes eligible[5] to perform the action at time s_u, she performs the action at time t_u, with $t_u \geq s_u$, 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 τ, a user receives badge b the first time she takes action a, potentially *increasing* its corresponding utility v_u. Here, the *delayed* introduction of badge b at time τ can be thought of as a natural experiment [25, 31], where the treatment of earning a badge is assigned *as if random* among users who are eligible to perform an action. Note that users who performed action a before the badge introduction time τ receive the badge automatically at τ, however, this does not influence our analysis since our focus is on the potential effect that first-time badges may have on the first time users take the action, and not on subsequent times.

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 present a high-level overview of our causal inference framework, highlighting the assumption it requires, and then elaborate on its individual components.

Our causal inference framework. Given an action of interest a, its corresponding first-time badge b with introduction time τ, the behavior of n users with respect to a, *i.e.*, $\mathcal{D}_a = \{(s_u, t_u, v_u)\}_{u \in [n]}$, one could just consider the set of users who performed the action for the first time before the badge introduction, *i.e.*, $t_u < \tau$, as control group, and the set of users who became eligible to perform the action after the badge introduction, *i.e.*, $\tau < s_u$, as treatment group. However, proceeding that way would have several limitations. First, a potential difference among the treatment and control groups could be due to random fluctuations in users' behavior over time due to many complex, confounding factors. Second, one would be unable to consider users who became eligible before the badge introduction but performed the action for the first time after the badge introduction. These users are the ones who were actually exposed to the badge introduction.

Our causal inference framework overcomes the above limitations by redefining the treatment and control groups as follows. We define the treatment group as the set of users who became eligible to perform the action at a time $s_u \in [\tau - w/2, \tau + w/2]$, where $w \geq 0$ is a given parameter. We define several control groups, each of them associated to a *virtual* badge b_i introduced at time $\tau_i \in [w/2, \tau - w] \cup [\tau + w, T - w/2]$, picked uniformly at random. Under these definitions, the treatment group represents the *true world* where the badge is introduced and the control groups represent different *counterfactual worlds* where a virtual badge, which models random fluctuation in users' behavior, is introduced. As long as the allocation of users across the treatment and control groups resemble

---

[5]Stack Overflow often requires a minimum reputation level for a user to be able to perform an action. For example, only users with at least a reputation level of 75 can offer a bounty.

random assignment, if the strength of the change induced by the badge in the true world (as measured by survival-based hypothesis testing) is larger than the strength of the changes induced by the virtual badges in a variety of counterfactual worlds, we can conclude that the change in the true world is not due to random fluctuation in users' behavior with higher confidence.

Next, we elaborate on two survival-based hypothesis testing procedures of increasing statistical power, which we use to measure the strength of the change induced by a (true or virtual) badge, and then describe how to formally compare the strength of the changes induced by the true and the virtual badges by means of a novel 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 [1]. Following the literature on temporal point processes, we represent such survival process as a binary counting process $N_u(t) \in \{0, 1\}$, 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., $\mathbb{E}[dN_u(t)] = \lambda_u(t)dt$, which we define as follows:

$$\lambda_u(t) = \begin{cases} 0 & \text{if } t < s_u \\ \lambda_0 & \text{if } s_u \leq t < \tau \\ \lambda_1 & \text{otherwise} \end{cases} \quad (2)$$

where s_u is the time when user u becomes eligible to perform action a, τ is the time when the first-time badge is introduced, and λ_0 and λ_1 are parameters shared across all users, which depend on the (intangible) utility users obtain from taking the action for the first time before and after the badge introduction, respectively.

Under this model, the null hypothesis $\mathcal{H}_0$, i.e., the badge did not have an effect, corresponds to $\lambda_0 = \lambda_1 \geq 0$ and the alternative hypothesis $\mathcal{H}_1$ 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:

$$\hat{\lambda}_0 = \hat{\lambda}_1 = \frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq T)}{\sum_{u \in [n]}(\min(t_u, T) - s_u)}, \quad (3)$$

and under the alternative hypothesis, they are given by:

$$\hat{\lambda}_0 = \frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq \tau)}{\sum_{u \in [n]}(\min(t_u, \tau) - s_u)\mathbb{I}(s_u < \tau)}$$
$$\hat{\lambda}_1 = \frac{\sum_{u \in [n]} \mathbb{I}(\tau < t_u \leq T)}{\sum_{u \in [n]}(\min(t_u, T) - \tau)\mathbb{I}(t_u > \tau)}, \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function and all the sums are over eligible users. Then, we can use a test statistic such as standard log-likelihood ratio (*LLR*) [18] to measure the strength of the change induced by the badge, i.e.,

$$LLR = \sum_{u \in [n]: t_u < T} \log f(s_u, t_u; \hat{\lambda}_0, \hat{\lambda}_1) - \log f(s_u, t_u; \hat{\lambda}_0)$$
$$+ \sum_{u \in [n]: t_u = \infty} \log S(s_u, T; \hat{\lambda}_0, \hat{\lambda}_1) - \log S(s_u, T; \hat{\lambda}_0)$$

where $f(s_u, t_u; \hat{\lambda}_0)$ and $f(s_u, t_u; \hat{\lambda}_0, \hat{\lambda}_1)$ are the likelihoods of the action times under the null and alternative model, respectively, and $S(s_u, T; \hat{\lambda}_0, \hat{\lambda}_1)$ and $S(s_u, T; \hat{\lambda}_0)$ are the corresponding survival functions. Here, the likelihoods and survivals can be computed using the intensity $\lambda_u(t)$ defined by Eq. 2, with $\lambda_0 = \lambda_1$ for the null model and $\lambda_0 \neq \lambda_1$ for the alternative model, as noted elsewhere [1].

— *Robust survival-based hypothesis testing:* The basic survival-based hypothesis testing procedure described above assumes the model parameters are shared across all users and, by doing so, it ignores the utility heterogeneity across users. Here, we account for the utility heterogeneity by considering different latent parameters per user, but sampled from the same distributions, i.e.,

$$\lambda_u(t) = \begin{cases} 0 & \text{if } t < s_u \\ \lambda_0(u) & \text{if } s_u \leq t < \tau \\ \lambda_1(u) & \text{otherwise} \end{cases}$$
$$\lambda_0(u) \sim \text{Gamma}(k_0, r)$$
$$\lambda_1(u) \sim \text{Gamma}(k_1, r) \quad (5)$$

where s_u is the time when user u becomes eligible to perform action a, τ is the time when the first-time badge is introduced, k_0, k_1 are shape parameters and r is a rate parameter. Here, note that $\mathbb{E}[\lambda_0] = k_0/r$ and $\mathbb{E}[\lambda_1] = k_1/r$.

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

$$\mathcal{H}_0 : k_0 = k_1 \geq 0$$
$$\mathcal{H}_1 : k_0 \neq k_1, k_0 \geq 0, k_1 \geq 0$$

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:

$$\hat{k}_0 = \hat{k}_1 = -\frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq T)}{\sum_{u \in [n]} \log\left(\frac{r}{r + \min(t_u, T) - s_u}\right)} \quad (6)$$

and under the alternative hypothesis, they are given by:

$$\hat{k}_0 = -\frac{\sum_{u \in [n]} \mathbb{I}(t_u \leq \tau)}{\sum_{u \in [n]} \log\left(\frac{r}{r + \min(t_u, \tau) - s_u}\right)\mathbb{I}(s_u \leq \tau)}$$
$$\hat{k}_1 = -\frac{\sum_{u \in [n]} \mathbb{I}(\tau < t_u \leq T)}{\sum_{u \in [n]} \log\left(\frac{r}{r + \min(t_u, T) - \tau}\right)\mathbb{I}(t_u > \tau)} \quad (7)$$

Similarly as in the basic survival model, we can then use a test statistic such as standard log-likelihood ratio (*LLR*) to measure the strength of the change induced by the badge.

— *Bootstrap difference-in-differences method:* Given an action of interest a, its corresponding first-time badge b with introduction time τ, the behavior of n users with respect to a, i.e., $\mathcal{D}_a = \{(s_u, t_u, v_u)\}_{u \in [n]}$, and a model-based hypothesis testing procedure as the ones described above, we assess the significance of the change induced by the true badge using the following bootstrap difference-in-difference method:
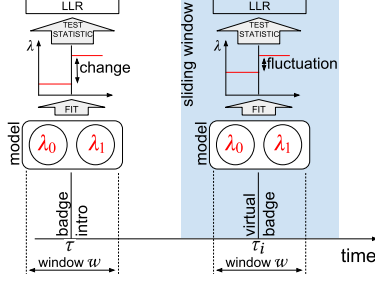
Figure 2: Our bootstrap difference-in-difference method. The treatment group (left) consists of users whose start time s_u lies in a window of size w around the time τ the badge is introduced. Each control group i (right) consists of users whose start time s_u lies in a window of size w around the time τ_i a virtual badge is introduced. The method runs hypothesis testing on both the treatment group and all control groups and then compares the test statistic (*e.g.*, LLR) of the treatment group with the empirical distribution of the test statistic of the control groups.

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, as treatment group. Then, we compute the maximum likelihood estimators of the parameters of the model of choice using Eqs. 3-4 (basic) or Eqs. 6-7 (robust) on that subset of users and use these estimated parameters to compute the log-likelihood ratio, *LLR*.

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 $s_u \in [\tau_i - w/2, \tau_i + w/2]$ as control group i. Finally, we compute the maximum likelihood estimators of the parameters of the model of choice, as in the previous step, on each of these subsets of users and use these estimated parameters to compute the log-likelihood ratio for each control group, LLR_{τ_i}.

III. We measure the strength of the change induced by the badge by means of the probability that the LLR of the control groups, for which the null hypothesis holds by design since virtual badges do not exist, is larger than the LLR of the treatment group, $p := F_{LLR}(LLR_\tau)$.

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 LLR under the null hypothesis $F_{LLR}(LLR)$ and a p-value $p = F_{LLR}(LLR_\tau)$, which accounts for the temporal fluctuations in users' behavior and allows us to reject the null hypothesis with higher confidence. Finally, as discussed previously, the main assumption needed for the above method to be valid is that the allocation of users across the treatment and control groups resemble random assignment. Table 2 shows this assumption is satisfied in our dataset, as we will discuss later.

| Average p-value | Rejection probability |
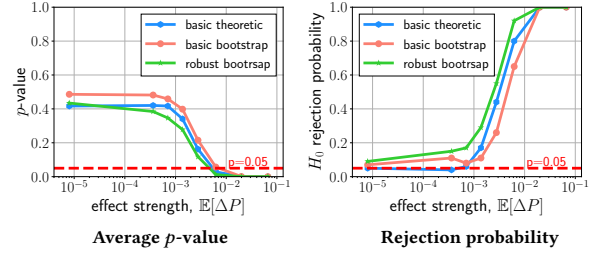
Figure 3: Performance of our causal inference framework on synthetic data. The left panel shows the average p-value against expected effect strength $\mathbb{E}[\Delta P]$, where lower (higher) is better for $\mathbb{E}[\Delta P] > 0$ ($\mathbb{E}[\Delta P] \approx 0$). The right panel informs about tests' statistical power by showing the rejection probability of the null hypothesis $\mathcal{H}_0$ at $p = 0.05$ against effect strength $\mathbb{E}[\Delta P]$, where higher (lower) is better for $\mathbb{E}[\Delta P] > 0$ ($\mathbb{E}[\Delta P] \approx 0$).

Framework evaluation. In this section, we compare the effectiveness of the basic survival model with the theoretical distribution of the LLR under the null hypothesis (χ_1^2, given by Wilks' theorem [34]; "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 $s_u \sim U[0, T]$, and her action time t from an intensity $\lambda_u(t)(1 + at)$, where $\lambda_u(t)$ is given by Eq. 5 and $a = 0.001$. Note that the term $(1 + at)$ imposes a global linear trend, which is often observed in real data[6]. Moreover, in Eq. 5, we set the badge introduction time to $\tau = T/2$, the rate parameter to $r = 10$, and consider different badge strength values expressed as the expected increase in the probability that users perform the action in less than $t = 10$ after s_u, which we denote as $\mathbb{E}[\Delta P]$. 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 $\mathcal{H}_0$ 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 $\mathcal{H}_0$ when a badge is introduced (*i.e.*, $\mathbb{E}[\Delta P] > 0$) while it is equally likely not to reject $\mathcal{H}_0$ when a badge is not introduced (*i.e.*, $\mathbb{E}[\Delta P] \approx 0$).

Remarks. We acknowledge that our framework has several limitations. In its current form, the framework is only applicable to first-time badges. In principle, it may be possible to augment our framework to other types of badges. For example, in threshold badges, one could replace the binary counting

---

[6]We obtain quantitatively similar results in the absence of a linear trend, *i.e.*, $a = 0$.

processes by non binary ones representing the number of actions $N(t)$ (*i.e.*, progress) and choose a functional form for the intensities depending on the numbers of actions, *i.e.*, $\lambda_1(N(t))$, as suggested in previous work [4]. However, one would need to address additional challenges, *e.g.*, accurate intensity models, and it is out of the scope of this work. Moreover, while our bootstrap difference-in-difference method does control for temporal fluctuations in the users' behavior, it cannot rule out the possibility that the change in users' behavior in the treatment group, if deemed significant, may be due to a hidden confounding factor rather than the badge introduction.

4 DO BADGES WORK?

Before we apply our causal inference framework to the three first-time badges described in Section 2, we first check the random assignment assumption between the treatment and control groups, which is necessary for our framework to provide sound conclusions. If the assignment is random, we would expect the groups (respective subpopulations) to be indistinguishable based on any additional covariates. Two groups are called indistinguishable (or balanced) if these additional covariates are within a standardized mean difference (SMD)[7] of 0.25 standard deviations [33]. In particular, Table 2 shows that treatment and control groups are indistinguishable in our dataset. Here, we would like to acknowledge that conclusions drawn from standardized mean differences (SMD) when evaluating variables with skewed distributions, as ours, should be taken with caution.

Once we have validated the random assignment assumption, we apply our causal framework to the three first-time badges. Figure 4 summarizes the results by means of:

(i) Test statistic over time for the basic and robust survival models, *i.e.*, LLR_{τ_i} and LLR_τ against τ_i and τ.

(ii) Empirical distribution of the test statistic under $\mathcal{H}_0$ 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 τ and $\hat{k}_1$ after τ, 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.

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 ~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

Table 2: Absolute Standardized Mean Difference (|SMD|) between the treatment and the control groups.

feature	Tag Editor		Prom. & Inv.	
	$|\widetilde{\text{SMD}}|$	$\sigma_{|\text{SMD}|}$	$|\widetilde{\text{SMD}}|$	$\sigma_{|\text{SMD}|}$
age	0.10	0.04	0.14	0.10
age-NA*	0.08	0.12	0.17	0.11
reputation	0.21	0.35	0.14	0.11
#views	0.16	0.23	0.11	0.08
#upvotes	0.18	0.28	0.13	0.10
#downvotes	0.08	0.07	0.03	0.02

*Balance check on a binary variable indicating if the value is missing.

than the intensities corresponding to the `Tag editor` or the `Investor` badge.

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 $\hat{k}_0 \leq 2 \cdot 10^{-4}$ to $\hat{k}_1 \approx 8 \cdot 10^{-4}$ for `Tag editor` badge and from $\hat{k}_0 \leq 5 \cdot 10^{-5}$ to $\hat{k}_1 \approx 2 \cdot 10^{-4}$ 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 (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[8]. Such discussions led to an increase on the minimum reputation one can transfer when offering a bounty.
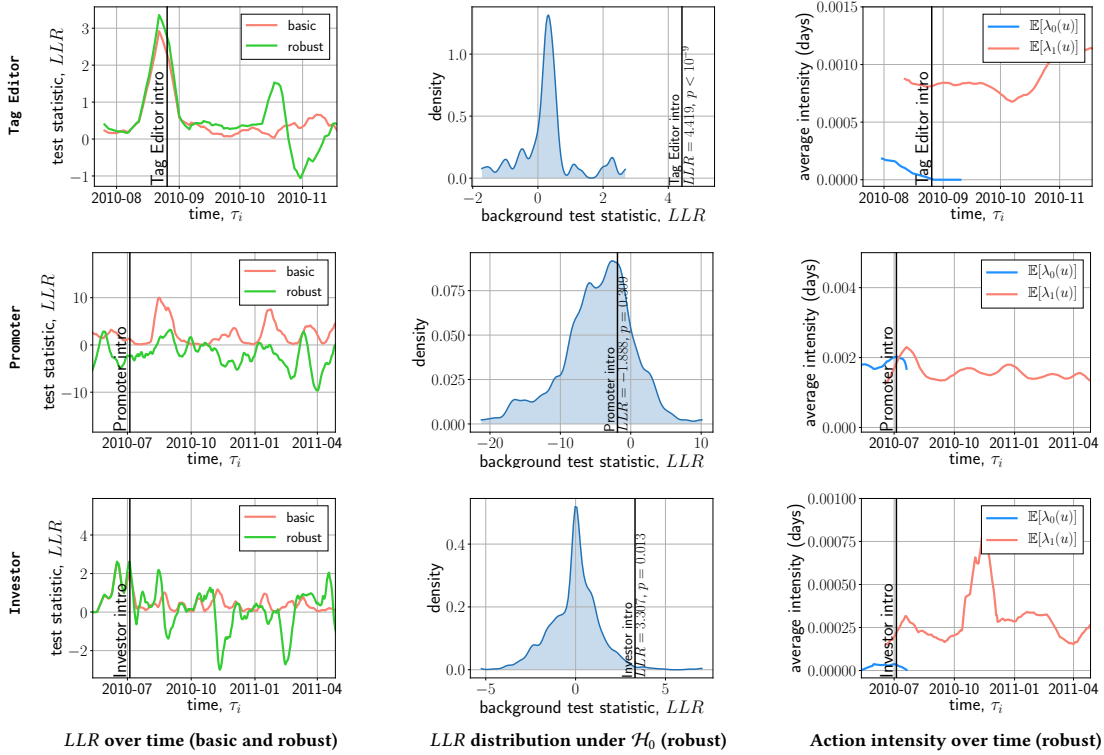
5 BADGES AND UTILITIES

In this section, we look for plausible reasons that explain *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.

User utilities. In the game theory literature [12, 19, 35], 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 to perform an action—is often modeled via a reservation value ω that the utility v must exceed in order for the user to perform the action. More specifically, if $p - c < w$, 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 ω, 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

---

[7]The standardized mean difference (SMD) is defined as the difference in means of the treatment and control group divided by the pooled standard deviation of both groups.

Figure 4: Causal effect of three badges (Tag Editor, Promoter and Investor) on users' behavior. Panels in the left show the test statistic (log-likelihood ratio; *LLR*) over time for the basic and robust survival models, *i.e.*, LLR_{τ_i} and LLR_τ against τ_i and τ. Panels in the middle show the empirical distribution of the test statistic (*LLR*) under the null hypothesis $\mathcal{H}_0$ and p-value for the robust survival model, *i.e.*, $F_{LLR}(LLR)$ and $p = F_{LLR}(LLR_\tau)$. Panels in the right show the average intensities for first-time action under robust survival model, *i.e.*, $\hat{k}_0$ before τ and $\hat{k}_1$ after τ, using a sliding window of length $w = 60$ days. The results suggest that the Tag editor and Investor badges were *successful* and the Promoter badge was *unsuccessful*.

explicitly adopt the above model and instead used a data-driven approach based on survival analysis using only the observable temporal traces. In this section, however, we turn our attention towards the above stylized model, try to identify measurable proxies of the model parameters for each of the studied badges and actions and use them to investigate the plausible 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 another user's question may obtain

less utility from the answers since they did not originally ask the question. Here, note that we did not find a correlation between the bounty value—its cost—and the final number of answers ($\rho \approx 0.1$) and thus we can safely ignore the bounty value.

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:

$$\lambda_p(t) = \begin{cases} 0 & \text{if } t < s \\ \lambda_0(p) & \text{if } s \le t < \tau \\ \lambda_1(p) & \text{otherwise} \end{cases} \tag{8}$$

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 τ is the time when the Tag Editor badge is introduced. Then, by comparing the maximum likelihood estimators of the model parameters, $\hat{\lambda}_0(p)$ and $\hat{\lambda}_1(p)$, for different
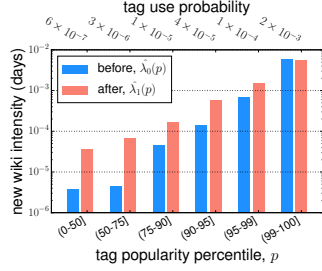
Figure 5: Causal effect of the `Tag Editor` badge for tags with different utility value, as estimated by their popularity level.

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. For example, while the intensity of the tags at the bottom 50% in terms of popularity increased an order of magnitude after the badge was introduced, the intensity of the tags at the top 1% did not change. This suggests that 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 (top panel); (ii) the introduction of the `Promoter` badge changed the users' willingness to offer bounties to their own questions only up to a very small degree (top panel, top figure), whereas the `Investor` badge did change it remarkably for bounties offered to other users' questions (bottom panel, bottom figure). In both cases, the change is statistically significant, *i.e.*, $\chi^2 = 86.2$, $p < 0.001$ for `Promoter` and $\chi^2 = 114.9$, $p < 0.001$ for `Investor` badge using Mood's median test, however, only the latter was sufficiently pointed to result in a significant change at an individual user level, as concluded in Section 4.

Remarks. For the specific actions and badges under studied, we were able to manually identify sensible proxies of the users' utilities, however, finding sensible proxies for other types of actions in Stack Overflow or other online platforms will require careful reasoning and justification.
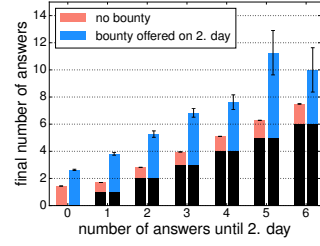
Total number of answers vs number of answers preceding bounty offering
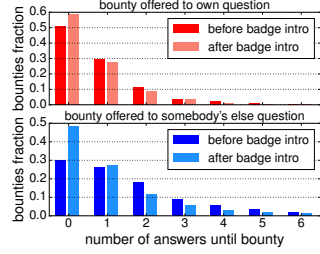
Number of answers preceding bounty offering

Figure 6: Causal effect of the `Promoter` and `Investor` badges for bounties with different utility value, as estimated by the number of answers preceding the bounty offering.

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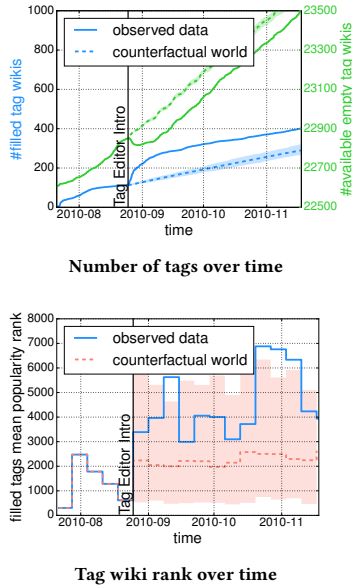
Counterfactual analysis. For the `Tag editor`, we assess the site functioning at a community level in terms of the number of new tag wikis over time and, for the `Investor` badges, we use the time to bounty and time to first answer across questions. More specifically, we proceed as follows.

— *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. 8 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. 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:

$$\lambda_b(t) = \begin{cases} 0 & \text{if } t < s \\ \lambda_0(b) & \text{if } s \leq t < \tau \\ \lambda_1(b) & \text{otherwise} \end{cases} \quad (9)$$

Number of tags over time

Tag wiki rank over time

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

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[9], and τ 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. 9, 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 top panel shows the number of unique tags with (blue) and without (green) supporting tag wiki over time in the true and in the simulated counterfactual world. Note that the number of unique tags without tag wiki grows over time as users ask questions with new (not previously used) tags. The results show that the tags with supporting tag wikis increase at a higher rate in the true world than in the counterfactual world just after the badge introduction, however, after a relatively short period of time, the rate decreased to its original value and match the rate at which tags with tag wikis grows in the counterfactual world. The bottom panel, which shows a decrease on the average rank popularity (rank) of the tags with supporting tag wikis, supports the following hypothesis to explain this phenomenon. The badge lifted the utility value of editing tag wikis, however, this increase was large enough to

---

[9]A bounty can only be offered two days after the question has been asked.
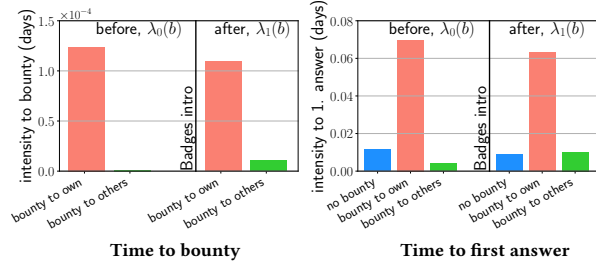
Figure 8: Time to bounty and first answer with and without the `Promoter` and `Investor` badges.

exceed the reservation value only 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. The latter observation suggests that the `Investor` badge may have mitigated the global slowdown in the time users take to answer questions by motivating more people to offer bounties faster—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 interesting to use our framework to analyze badges in other online platforms as well as extend it 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.

REFERENCES

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

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

[3] Tim Althoff, Pranav Jindal, and Jure Leskovec. 2017. 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 (WSDM 2017)*. 537–546.

[4] Ashton Anderson, Daniel P. Huttenlocher, Jon M. Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd International World Wide Web Conference (WWW '13)*. 95–106.

[5] Ashton Anderson, Daniel P. Huttenlocher, Jon M. Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd International World Wide Web Conference (WWW '14)*. 687–698.

[6] Sinan Aral and Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* 337, 6092 (2012), 337–341.

[7] Benny Bornfeld and Sheizaf Rafaeli. 2017. Gamifying with badges: A big data natural experiment on Stack Exchange. *First Monday* 22, 6 (2017).

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

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

[10] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. Gamification. using game-design elements in non-gaming contexts. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI 2011)*. 2425–2428.

[11] Stephen G. Donald and Kevin Lang. 2007. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics* 89, 2 (2007), 221–233.

[12] David Easley and Arpita Ghosh. 2016. Incentives, Gamification, and Game Theory: An Economic Approach to Badge Design. *ACM Trans. Economics and Comput.* 4, 3 (2016), 16:1–16:26.

[13] Oriol Borrás Gené, Margarita Martínez-Núñez, and Ángel Fidalgo Blanco. 2014. Gamification in MOOC: challenges, opportunities and proposals for advancing MOOC model. In *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM '14)*. 215–220.

[14] Stefan Göbel, Sandro Hardy, Viktor Wendel, Florian Mehm, and Ralf Steinmetz. 2010. Serious games for health: personalized exergames. In *Proceedings of the 18th International Conference on Multimedia 2010 (MM '10)*. 1663–1666.

[15] Juho Hamari, David J. Shernoff, Elizabeth Rowe, Brianno Coller, Jodi Asbell-Clarke, and Teon Edwards. 2016. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior* 54 (2016), 170–179.

[16] Michael D. Hanus and Jesse Fox. 2015. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education* 80 (2015), 152–161.

[17] Philipp Herzig, Susanne Strahringer, and Michael Ameling. 2012. Gamification of ERP systems - Exploring gamification effects on user acceptance constructs. In *Multikonferenz Wirtschaftsinformatik*. 793–804.

[18] Robert V. Hogg and Allen T. Craig. 1995. *Introduction to mathematical statistics*. Prentice Hall.

[19] Nicole Immorlica, Gregory Stoddard, and Vasilis Syrgkanis. 2015. Social Status and Badge Design. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. 473–483.

[20] Abigail Z. Jacobs, Samuel F. Way, Johan Ugander, and Aaron Clauset. 2015. Assembling thefacebook: Using Heterogeneity to Understand Online Social Network Assembly. In *Proceedings of the 7th ACM Web Science Conference (WebSci '15)*. 18:1–18:10.

[21] Daniel Johnson, Sebastian Deterding, Kerri-Ann Kuhn, Aleksandra Staneva, Stoyan Stoyanov, and Leanne Hides. 2016. Gamification for health and wellbeing: A systematic review of the literature. *Internet interventions* 6 (2016), 89–106.

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

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

[24] Simon McCallum and Costas Boletsis. 2013. Dementia Games: A Literature Review of Dementia-Related Serious Games. In *Proceedings of the 4th International Conference on Serious Games Development and Applications (SGDA 2013)*. 15–27.

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

[26] Benedikt Morschheuser, Juho Hamari, and Jonna Koivisto. 2016. Gamification in Crowdsourcing: A Review. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS 2016)*. 4375–4384.

[27] Cristina Muntean. 2011. Raising engagement in e-learning through gamification. In *Proceedings 6th International Conference on Virtual Learning (ICVL 2011)*.

[28] Tobias Mutter and Dennis Kundisch. 2014. Behavioral Mechanisms Prompted by Badges: The Goal-Gradient Hypothesis. In *Proceedings of the International Conference on Information Systems - Building a Better World through Information Systems (ICIS 2014)*.

[29] Hüseyin Oktay, Brian J. Taylor, and David D. Jensen. 2010. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis (SNAKDD 2009)*. 1–9.

[30] Tuan Q. Phan and Edoardo M. Airoldi. 2015. A natural experiment of social network formation and dynamics. *Proceedings of the National Academy of Sciences* 112, 21 (2015), 6595–6600.

[31] Mark R. Rosenzweig and Kenneth I. Wolpin. 2000. Natural "natural experiments" in economics. *Journal of Economic Literature* 38, 4 (2000), 827–874.

[32] Katie Seaborn and Deborah I. Fels. 2015. Gamification in theory and action: A survey. *Int. J. Hum.-Comput. Stud.* 74 (2015), 14–31.

[33] Elizabeth A. Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1–21.

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

[35] Jiawei Zhang, Xiangnan Kong, and Philip S. Yu. 2016. Badge System Analysis and Design. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016)*.