On Validation and Predictability of Digital Badges' Influence on Individual Users

Tomasz Kuśmierczyk and **Kjetil Nørvåg**
Norwegian University of Science and Technology
{tomaszku,noervaag}@ntnu.no

Abstract

Badges are a common, and sometimes the only, method of incentivizing users to perform certain actions on online sites. However, due to many competing factors influencing user temporal dynamics, it is difficult to determine whether the badge had (or will have) the intended effect or not.

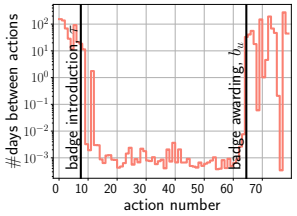
In this paper, we introduce two complementary approaches for determining badge influence on users. In the first one, we cluster users' temporal traces (represented with Poisson processes) and apply covariates (user features) to regularize results. In the second approach, we first classify users' temporal traces with a novel statistical framework, and then we refine the classification results with a semi-supervised clustering of covariates.

Outcomes obtained from an evaluation on synthetic datasets and experiments on two badges from a popular Q&A platform confirm that it is possible to validate, characterize and to some extent predict users affected by the badge.

Introduction

Awarding a digital badge after a user performs certain actions is a common mechanism to motivate users on online sites, be it social networking sites like Foursquare[1], education sites like Khan Academy[2], or crowdlearning Q&A sites like Stack Overflow[3]. Previously, there have been several attempts at modeling the badges' effect on online communities and at recommending how the badge systems should be designed. However, there are no previous studies actually verifying whether the badges have any impact on individual users or not; it has been taken for granted that badges affect targeted users in a desired way. In contrast, in this paper we take a closer look at this assumption, and present the first work that addresses the problems of *validation, characterization and prediction of users attracted to badges*.

It is a challenging task to answer the question whether a user was (is) in any way motivated by the badge. Users tend to evolve over time, and apart from badges there are usually many competing factors influencing their dynamics. Furthermore, neither ground truth nor counterfactual data showing behavior of users not influenced by badges are available.



user characteristics	
age	40
location	Madison
views	102981
comments	3246
upvotes	14407
downvotes	9019

Figure 1: Sample user from Stack Overflow influenced by the *Research Assistant* badge that is awarded for tag wiki edits. The user increases its action rate when the badge is introduced to the community (at time τ), and returns to the previous rate after receiving it (at time b_u).

In this paper, we focus on the most popular type of badges, *i.e.*, *threshold badges*, that are awarded after a user performs a certain number of desired actions. The above challenges in this context can be addressed by simultaneously looking at users' temporal traces and their general characteristics. In particular, we identified the following useful patterns:

- attracted users change their *mean* behavior around the badge awarding time
- influenceable users are similar

As an example, Figure 1 illustrates how user action rate changes due to the badge, and also shows the associated user features and statistics.

In this paper, we propose two complementary solutions to the *user badge influence problem* exploiting the above observations. In both we apply users' temporal traces (modeled as *non-homogeneous Poisson processes*) as a main source of the badge effect information and use associated covariates (user features and statistics) to regularize classification results.

This paper makes the following contributions:

- introduction and formalization of the *user badge influence problem*
- validation and prediction of the influence of badges on individual users with two novel methods:
  I. a model-based algorithm to cluster (problem-specific) counting processes with covariates used to encode priors

[1] https://foursquare.com
[2] https://www.khanacademy.org/
[3] https://stackoverflow.com

II. a statistical test complemented with a way of calibrating it by means of *virtual badges bootstrapping* and an adaptation of the EM clustering algorithm that refines the results of this test

– empirical evaluation using synthetic data

– case studies of two badges from a popular Q&A platform

The code used in this paper is publicly available [4].

In the rest of the paper we'll give an overview of related work, formalize the problem, provide a detailed description of our proposed solutions, and demonstrate their effectiveness on synthetic and real datasets.

Related Works

Our work extends previous studies on motivational mechanisms in social media (Ghosh and McAfee 2011; Hamari, Koivisto, and Sarsa 2014; Lewis 2004) and in particular on understanding and modeling the effects of badges (Anderson et al. 2013; Gibson et al. 2015; Zhang, Kong, and Yu 2016).

Previously, (Immorlica, Stoddard, and Syrgkanis 2015; Easley and Ghosh 2016; Zhang, Kong, and Yu 2016) worked on optimal badges design from a game-theoretic perspective. They relied on strong theoretical assumptions not necessarily satisfied in real data, including the assumption that badges *always* work. In this paper, we challenge this presumption. Early studies suggesting that may not always be the case appeared first in the educational context (Abramovich, Schunn, and Higashi 2013). In the context of social media the problem of badges effectiveness was noticed very recently (Bornfeld and Rafaeli 2017; Kusmierczyk and Gomez-Rodriguez 2017; Hamari 2017). In these works researchers assumed badges effect to be binary, *i.e.*, either a badge changes community functioning or not, and (apart from (Kusmierczyk and Gomez-Rodriguez 2017) who on the other hand worked with simpler *single-action badges*) focused on site-global statistics (total number of views, edits, etc.), whereas we take a user-level perspective and try to understand badge influence on individuals. This places our study closer to works trying to characterize susceptible users in social media, for example (Aral and Walker 2012), and modeling user behavior in presence of badges (Anderson et al. 2013; Mutter and Kundisch 2014). The latter two rely on a *goal-gradient hypothesis*, *i.e.*, users become more active closer to a badge. For example, Anderson et al. introduced a game-theoretic model in which the badge *always* confers a value to a user in each step after she receives it. Optimal user policy for that model is then a (mentioned above) goal-gradient behavior, that is then sought in empirical evaluation. However, the plots they present show results that were averaged over all the users, whereas we observe that many individuals diverge from this behavior. Therefore, in contrast to the previous works, we decided to take a data-driven approach where, apart from quantifying the badge effect, we also relax assumptions about the behavior of users, *e.g.*, we focus on *mean* changes in user behavior around the time of badge awarding. In consequence, we are able to study badge-related patterns on a more granular (*i.e.*, individual user) level.

[4] http://github.com/tkusmierczyk/badges2

Problem Formulation

In this section, we present the problem of *determination of badge influence on a user* in a formal way, and introduce a point process model of user behavior in context of a badge.

Notation: Badges and Users. A digital *threshold badge* b can be represented with a tuple:

$$b \;:=\; (\underbrace{\tau}_{\text{introduction}},\; \overbrace{a}^{\text{desired actions type}},\; \underbrace{T}_{\text{threshold}}),$$

where τ is the badge introduction time (= the time when the badge started being awarded), a is an assigned action type, and T is the badge threshold (= the number of type a actions that need to be performed by a user in order to be awarded the badge).

User $u \in U$ in context of the badge b can be represented by a tuple:

$$u \;:=\; (\underbrace{\vec{x_u}}_{\text{user features}},\; \overbrace{\{t_u\}}^{\text{action times}},\; \underbrace{s_u}_{\text{start time}},\; \overbrace{e_u}^{\text{end time}},\; \underbrace{b_u}_{\text{badge awarding}},\; \overbrace{i_u}^{\text{badge attraction}}),$$

where $\vec{x_u}$ is a vector of badge covariates (*e.g.*, user characteristics), $\{t_u\}$ is a set of timestamps of desired (=type a) actions, s_u and e_u designate user activeness interval (time span in which we test badge effect), b_u is the time when user u received badge b (=achieved level of T actions). If the user u has not received the badge yet (*i.e.*, $|\{t_u\}| < T$), we set $b_u = \infty$. Finally, the binary variable i_u informs if the user is/was attracted by the badge reward perspective or not (the fact that user received a badge does not necessary imply that she had any interest in that – it could be just a side-effect of her normal activity).

Additionally, to simplify some of the later formulations, we define: $l_u = e_u - s_u$, $l_u^0 = b_u - s_u$, $l_u^1 = e_u - b_u$, $n_u = |\{t_u\}|$, $n_u^0 = |\{t_u :\ t_u < b_u\}|$, $n_u^1 = |\{t_u :\ t_u \geq b_u\}|$.

Influenced Users Validation and Prediction. We distinguish users influenced by the badge b from those not attracted via the binary variable i_u. Unfortunately, the variable is usually hidden. Its value recovery can be done in two practical settings:

I. *Validation*: user received the badge ($b_u < \infty$) and we verify if it did not happen just by chance.

II. *Prediction*: user has not received the badge yet ($b_u = \infty$) and we try to forecast if she may be interested in receiving it.

For neither of the tasks we know the truth. Therefore, we rely only on our assumptions relating badge influence with temporal traces ($\{t_u\}$) and users' general characteristics (encoded in $\vec{x_u}$).

Temporal Traces Model. It might be hard to observe if users attracted by the badge change their behavior when they receive it. However, temporal fluctuations and impact of competing factors can be reduced with averaging over time. In particular, inspired by the survival model from the work by (Kusmierczyk and Gomez-Rodriguez 2017), we assume the following model of underlying temporal traces, where user u's action times are drawn from the *non-homogeneous*

Poisson process (Daley and Vere-Jones 2002) controlled by the intensity $\lambda_u(t)$ that takes one of the two forms: $\lambda_u^0(t)$ or $\lambda_u^1(t)$, depending on the latent variable i_u:

- $i_u = 0$ (user not attracted by the badge): intensity is a constant (user does not change her behavior over time)
- $i_u = 1$ (user attracted by the badge): actions mean intensity changes when the badge is awarded at b_u

Formally, the model is expressed as follows:

$$\begin{aligned} \{t_u\} &\sim PP(\lambda_u^{i_u}(t)) \\ \lambda_u^0(t) &= \begin{cases} 0 & \text{if } t < s_u \vee t > e_u \\ \lambda^0(u) & \text{otherwise} \end{cases} \\ \lambda_u^1(t) &= \begin{cases} 0 & \text{if } t < s_u \vee t > e_u \\ \lambda_0^1(u) & \text{if } s_u < t \leq b_u \\ \lambda_1^1(u) & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

Learning Attracted Users via Poisson Processes Clustering

In this section, we introduce a novel model-based algorithm to cluster Poisson processes, that we use to identify users influenced by the badge. Its extended version employs covariates to regularize clusters assignment priors and allows for new users prediction.

Basic Model. We assume that the fraction of users attracted by the badge (having $i_u = 1$) is π, and the intensities λ (expressed in Eq. 1) come from the shared prior gamma distributions:

$$\begin{aligned} i_u &\sim \text{Bernoulli}(\pi) \\ \lambda^0(u) &\sim \text{Gamma}(\alpha^0, \beta^0) \\ \lambda_0^1(u) &\sim \text{Gamma}(\alpha_0^1, \beta_0^1) \\ \lambda_1^1(u) &\sim \text{Gamma}(\alpha_1^1, \beta_1^1) \end{aligned} \tag{2}$$

The full model then has seven hyperparameters: $\theta^0 = \{\alpha^0, \beta^0\}$, $\theta^1 = \{\alpha_0^1, \beta_0^1, \alpha_1^1, \beta_1^1\}$ steering the behavior of users with respectively $i_u = 0$ and $i_u = 1$, and π controlling the fraction of users attracted by the badge. Latent variables are i_u and action intensities $\lambda^0(u)$ and $\{\lambda_0^1(u), \lambda_1^1(u)\}$.

The model can be factorized thanks to independence between user probabilities and independence between Poisson processes on non-overlapping intervals, and then simplified via marginalization of latent intensities. The procedure leads to the following conditional user probabilities:

$$\begin{aligned} P(\{t_u\}|\theta_0, i_u = 0) =& \frac{{\beta^0}^{\alpha^0}}{(l_u + \beta^0)^{\alpha^0 + n_u}} \frac{\Gamma(\alpha^0 + n_u)}{\Gamma(\alpha^0)} \\ P(\{t_u\}|\theta_1, i_u = 1) =& \frac{{\beta_0^1}^{\alpha_0^1}}{(l_u^0 + \beta_0^1)^{\alpha_0^1 + n_u^0}} \frac{\Gamma(\alpha_0^1 + n_u^0)}{\Gamma(\alpha_0^1)} \cdot \\ & \cdot \frac{{\beta_1^1}^{\alpha_1^1}}{(l_u^1 + \beta_1^1)^{\alpha_1^1 + n_u^1}} \frac{\Gamma(\alpha_1^1 + n_u^1)}{\Gamma(\alpha_1^1)} \end{aligned} \tag{3}$$

The model collapses to a mixture-model with two clusters determined by $i_u = 0$ and $i_u = 1$ and controlled by hyperparameters θ^0 and θ^1, and with mixing factor π. Cluster assignments and hyperparameters in this class of models are typically inferred with an *EM-like procedure* that consists of two alternating steps taking in our case the following form:

I. *Maximization*: hyperparameters are updated:

$$\begin{Bmatrix} \theta_{new}^0 \\ \theta_{new}^1 \\ \pi_{new} \end{Bmatrix} = \underset{\theta^0, \theta^1, \pi}{\text{argmax}} \sum_u \log P(\{t_u\}, i_u | \theta^0, \theta^1, \pi)$$

where the complete-data likelihood per user relies on per-cluster user likelihoods expressed in Eq. 3:

$$\begin{aligned} \log & P(\{t_u\}, i_u | \theta^0, \theta^1, \pi) = \\ & \gamma(i_u)(\log P(\{t_u\}|\theta^1, i_u = 1) + \log \pi) + \\ & (1 - \gamma(i_u))(\log P(\{t_u\}|\theta^0, i_u = 0) + \log(1 - \pi)) \end{aligned}$$

A closed-form solution to the optimization problem does not exist. Instead, we first find cluster probabilities:

$$\pi_{new} = \frac{\sum_{u \in U} \gamma(i_u)}{|U|} \tag{4}$$

and then resort to numerical optimization with positivity constraints to find θ_{new}^0 and θ_{new}^1

II. *Expectation*: posterior cluster responsibilities are found in the usual way:

$$\gamma(i_u) = \frac{P(\{t_u\}|\theta^1, i_u = 1)\pi}{P(\{t_u\}|\theta^1, i_u = 1)\pi + P(\{t_u\}|\theta^0, i_u = 0)(1 - \pi)}$$

Including Covariates. Badges attract users of similar characteristics and therefore user influence covariates $\vec{x_u}$ can be applied for clustering improvement as a form of regularization. We incorporate them in the our hierarchical model similar to (Liang et al. 2016) by replacing the constant cluster membership prior with user personalized ones, *i.e.*, $\pi \rightarrow \pi_u$ that we furthermore posit to have a functional form:

$$\pi_u = f(\vec{x_u}, \vec{w}) \in [0, 1] \tag{5}$$

where $\vec{w}$ are parameters of the function f. In general, f can be any function (for example neural network) but due to its simplicity we choose logistic regression, *i.e.*, $f(\vec{x_u}, \vec{w}) = \text{sigmoid}(\vec{w} \cdot \vec{x_u})$.

Conditional independence between the priors for cluster memberships and clusters' parameters imply that the inference procedure described above can be adjusted in a simple way by replacing the updates in Eq. 4 with the following optimization of vector $\vec{w}$:

$$\vec{w}_{new} = \underset{\vec{w}}{\text{argmax}} \sum_u (f(\vec{x_u}, \vec{w}) - \gamma(i_u))^2$$

Validation. In the proposed approach the variable i_u encoding user attraction towards the badge is also used to select between clusters. In particular, the posterior probabilities of cluster assignments can be interpreted as the badge impact probabilities. Ideally (*i.e.*, with probability 1), users with $i_u = 0$ should be assigned to the first cluster and with $i_u = 1$ to the second cluster.

Prediction. For a new user without temporal trace we predict badge attraction only relying on her features and statistics:

$$\hat{P}(i_u) = f(\vec{x_u}, \vec{w})$$

Learning Attracted Users with NHST and Covariates Clustering

In this section, we propose a two-phase procedure validating badge influence on users. In the first phase, we approximately identify users influenced by the badge with a robust *Null Hypothesis Significance Testing* (NHST) procedure. In the second phase, we refine assignments with a semi-supervised clustering of covariates.

Robust Validation of Attracted Users

Behavior Change Testing. The alternative that a user u was or was not attracted by the badge can be expressed in terms of the null and the alternative hypotheses:

$$H_0 \,:\, i_u = 0 \text{ (\textit{badge b did not have an effect on user u})}$$
$$H_1 \,:\, i_u = 1 \text{ (\textit{badge b influenced user u})}$$

Under the model in Eq 1 we can restate it in the following way:

$$H_0 \,:\, \lambda_u^{i_u} = \lambda_u^0(t)$$
$$H_1 \,:\, \lambda_u^{i_u} = \lambda_u^1(t)$$

Test Statistic. We use a standard log-likelihood ratio between likelihoods corresponding to H_0 and H_1 as a test statistic (Hogg and Craig 1995) which in our case takes the following form:

$$LLR(\lambda^0(u), \lambda_0^1(u), \lambda_1^1(u)) = n_u \log \lambda^0(u) - l_u \lambda^0(u) + \\ -n_u^0 \log \lambda_0^1(u) + l_u^0 \lambda_0^1(u) - n_u^1 \log \lambda_1^1(u) + l_u^1 \lambda_1^1(u)$$

where we plug-in MLE estimates for respective intensities: $\hat{\lambda^0}(u) = \frac{n_u}{l_u}$, $\hat{\lambda_0^1}(u) = \frac{n_u^0}{l_u^0}$, $\hat{\lambda_1^1}(u) = \frac{n_u^1}{l_u^1}$ and assume that $(0 \log 0) = 0$.

Robust Estimation of the Test Statistic Distribution. Asymptotically the test statistic $-2LLR$ for *nested models* has an approximate chi-square distribution (Wilks 1938) with the number of degrees of freedom equal to difference between compared models, *e.g.*, in our case $df = 1$. The test statistic transformation to p-value is then given by: $p \approx 1 - \boldsymbol{\chi}^2_{df}(-2LLR)$ where $\boldsymbol{\chi}^2$ is a chi-square CDF.

The standard procedure can detect a change in user behavior happening around b_u, but is not able to differentiate between the badge causal effect and other competing factors. Instead, we design and apply the calibration procedure (similar to (Kusmierczyk and Gomez-Rodriguez 2017)) that accounts for them by simulating a counter-factual world where the badge was never awarded and measuring the strength of observed changes there. In practice the test statistic empirical distribution is estimated with the following *virtual badges bootstrapping* procedure:

1. Sample B *virtual badges* $b'_u \sim U([s_u, b_u - m] \cup [b_u + m, e_u])$ where m is some small margin.
2. Remove the true badge effect by putting it outside the updated activeness limits: $(s'_u,\ e'_u) \,=\, \begin{cases} (s_u, b_u - m) & \text{if } b'_u < b_u \\ (b_u + m, e_u) & otherwise \end{cases}$
3. Evaluate LLR' with simulated b'_u, s'_u, e'_u and adequately updated $\{t'_u\}$.
4. Approximate empirical p-value: $p \,=\, \frac{|\{LLR' > LLR\}|}{B}$

Assignment Refining via Semi-supervised Clustering of Covariates

NHST Assignments Misclassification. The above testing procedure applied to each user splits the population into two groups: positives P for whom we managed to reject H_0 at significance level α and negatives N for whom we failed to reject H_0. Although this can be used as a first approximation to i_u, both groups contain many misclassified cases. In particular, the *false positives rate* (*FPR*) and the *false negatives rate* (*FNR*) depend on the *statistical test power* and *prevalence* of the positives over negatives, that both are unknown. For example, (Sellke, Bayarri, and Berger 2001; Colquhoun 2014) estimate *FPR* to be at least around 25% when the prior probability of a real effect is 0.5 and $\alpha = 0.05$. This means that at least $1/4$ of users initially assigned $i_u = 1$ actually have $i_u = 0$. For $i_u = 0$, the fraction of misclassified cases would be even higher.

Semi-Supervised Clustering with Group Priors. We achieve the reduction of the above classification error employing a novel semi-supervised extension to the standard *EM algorithm for gaussian mixtures* (Bishop 2006) The extended algorithm works (=clusters users) in covariates space but additionally employs the information transferred from the first (=NHST) phase. In particular, initial user assignments and our beliefs about misclassification rates we encode in priors to cluster assignments (=mixing coefficients).

NHST classification splits users into two groups, where P and N are respectively users initially classified as positives and negatives. For each group $G \in \{P, N\}$ we propose to use separate mixing coefficients $\vec{\pi_G}$ with Dirichlet hyperpriors, *i.e.*, $\vec{\pi_G} \sim \text{Dirichlet}(\alpha_G^0, ..., \alpha_G^K)$, where K is a standard parameter controlling the number of clusters (in contrast to Poisson processes clustering, in covariates space we can have arbitrary number of clusters). In order to be able to interpret clustering results, for each cluster we assign either $i_u = 1$ (clusters denoted as C^1) or $i_u = 0$ (clusters denoted as C^0). Finally, we can initialize the algorithm as follows:

$$\alpha_G^c \,=\, \begin{cases} \sigma \frac{|P| \cdot \textit{FPR}}{|C^0|} & \text{if } c \in C^0 \wedge G = P \\ \sigma \frac{|P| \cdot (1 - \textit{FPR})}{|C^0|} & \text{if } c \in C^1 \wedge G = P \\ \sigma \frac{|N| \cdot (1 - \textit{FNR})}{|C^1|} & \text{if } c \in C^0 \wedge G = N \\ \sigma \frac{|N| \cdot \textit{FNR}}{|C^1|} & \text{if } c \in C^1 \wedge G = N \end{cases}$$

Values of α_G^c encode beliefs of how many users from group G should end up in cluster c according to our trust in the initial classification based on NHST. Parameter σ balances between classification and clustering impact and informs how sure we are about the values of *FPR* and *FNR*. For example, we use $\textit{FPR} = 0.25$, $\textit{FNR} = 0.4$ and $\sigma = 1.0$.

The model fitting is performed in a standard way via EM, apart from two differences: (1) when calculating expectations new priors $\vec{\pi_G}$ are used, and (2) in the maximization step π_G^c are updated per group: $\pi_G^c \sim \alpha_G^c + \sum_{u \in G} \gamma(z_u^c)$, where $\gamma(z_u^c)$ are posterior cluster responsibilities.

Validation. The above procedure results in assigning each user a vector of cluster probabilities. The probability of the badge influencing a user can be then calculated as a total probability of clusters with $i_u = 1$.

Prediction. Prediction of new users can be performed via co-clustering. Specifically, users for which we could not perform the statistical test we include into the clustering as a new group X with uninformative priors, for example $\alpha_X^c = 1$. The rest of the method remains unaltered.

Co-clustering of users with badge and without badge can improve classification results in both validation and prediction, but the data distributions must be similar in terms of grouping attracted and not-attracted users. This assumption may be hard to ensure for real data. Therefore, to improve robustness and prediction quality, we propose to first cluster users with badge using the above procedure and then employ clustering results to train a standard classifier with better generalization properties, for example logistic regression.

Synthetic Data Evaluation

In this section, we compare with the help of synthetically generated data the effectiveness of the proposed approaches for validation and prediction of users attracted by the badge.

Basic Setting. We simulate the behavior of $N = 1000$ users: $N/2$ users with both temporal dynamics $\{t_u\}$ and covariates $\vec{x_u}$, and $N/2$ users with only covariates $\vec{x_u}$, that imitate new users. For each user we assign a latent variable i_u: with probability π: $i_u = 1$ and with probability $1 - \pi$: $i_u = 0$.

The users' temporal traces $\{t_u\}$ we sample according to intensities expressed in Eq. 1. The intensities $\lambda^0(u)$ and $\lambda_1^1(u)$ we draw according to Eq. 2 where we fix variances $Var(\lambda^0(u)) = Var(\lambda_1^1(u)) = 25$ and means $E(\lambda^0(u)) = 10$, $E(\lambda_1^1(u)) = 10 - \Delta_\lambda$, and intensity $\lambda_0^1(u)$ we fix respectively to $\lambda_0^1(u) = \lambda_1^1(u) + 2\Delta_\lambda$. The parameter Δ_λ *controls the strength of the simulated badge effect*. In the basic setting, randomness in individual user temporal trace $\{t_u\}$ appears due to point processes sampling procedure.

Users are independent, and therefore without loss of generality we can assume start times for all users $s_u = 0$. Furthermore, we set badge awarding time to $b_u = 100/\lambda^0(u)$ for users with $i_u = 0$, and $b_u = 100/\lambda_0^1(u)$ for users with $i_u = 1$. User end times we set to $e_u = b_u + u \cdot b_u$ ($u \sim U[0, 1]$).

We sample user features from bivariate (=two features per user) normal distributions:

$$\vec{x_u} \sim \begin{cases} N(0, \Sigma) & \text{when } i_u = 0 \\ N(\Delta_x \cdot s_{max}\vec{v}_{max}, \Sigma) & \text{otherwise} \end{cases}$$

where covariance matrix

$$\Sigma \sim \text{Wishart}\,(10, \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}),$$

s_{max} is the largest singular value of Σ corresponding to eigenvector $\vec{v}_{max}$, and Δ_x *controls discrepancy between features of users from different groups*.

Disturbed Data Setting. We study the robustness of our methods by simulating disturbed data, *e.g.*, temporal fluctuations in user intensities. In particular, we add (typical for real data) temporal trend, *i.e.*, $\lambda_u^{i_u}(t)$ becomes $\lambda_u^{i_u}(t)(1 + At)$, where A controls the extent of the simulated fluctuation.

Evaluation. In contrast to what is the case for real data, for synthetic data we know user attitude towards the badge (i_u) and therefore we can evaluate prediction results against it. Specifically, we employ *Area Under Curve* (AUC) that accounts for uncertainty in our methods predictions. We measure AUC separately for users with full information (user validation problem) and for users with limited data (user prediction problem). Every experiment we repeat 20 times and then average results.

Results. Figure 2 summarizes the simulation results in the basic setting for varying badge effects (Δ_λ) and clusterization levels (Δ_x). The methods based on our *2-phase* procedure (we show results only for *bootstrap* variant; results for *theoretic* variant in the basic setting are identical) improve over the basic *NHST* classification and have a superior performance over *Poisson (processes) clustering*. Poisson clustering fails when intensity differences between user classes are small (*e.g.*, $\Delta_\lambda \leq 0.5$). We found that the method in the considered variant (with logistic regression in Eq. 5) does not benefit sufficiently from differences between users' covariates (Δ_x) and as a result underestimates the probabilities of badges having effect.

Fluctuations in the temporal data lead to performance degradation. For example, Figure 3 shows how AUC is decreasing when divergence from the models (controlled by trend A) is increasing. The most affected method is *2-phase theoretic* (2-phase procedure with theoretic estimation of the test statistic distribution). The intermediate results (*NHST theoretic*) confirm that the method tends to overestimate badge influence when fluctuations are strong. However, the least affected is *Poisson processes clustering*, it never outperforms *2-phase bootstrap* - the robust variant of our 2-phase procedure.

Finally, we investigate our methods in terms of sensitivity to class imbalance (Figure 4). Class imbalance has a low impact on validation performance. It happens because the main indicator of the badge influence on user is a change in individual user dynamics around the badge awarding time and covariates are only 'regularizers'. On the other hand, the prediction relies entirely on covariates and if one class is underrepresented covariates distributions are poorly fitted and prediction fails, *e.g.*, AUC approaches 0.5 for both very small and large π.

Real Data Experiments

In this section, we investigate the effectiveness of our methods when applied to real data, *i.e.*, two sample badges from a popular Q&A platform.

Data Description and Preprocessing. In the real-data experiments we used a Stack Overflow dataset [5], that contains timestamped events from between July 2008 and September 2014 and some basic information (profile and actions record) about registered users. In particular, as badge *covariates* we

[5] https://archive.org/details/stackexchange

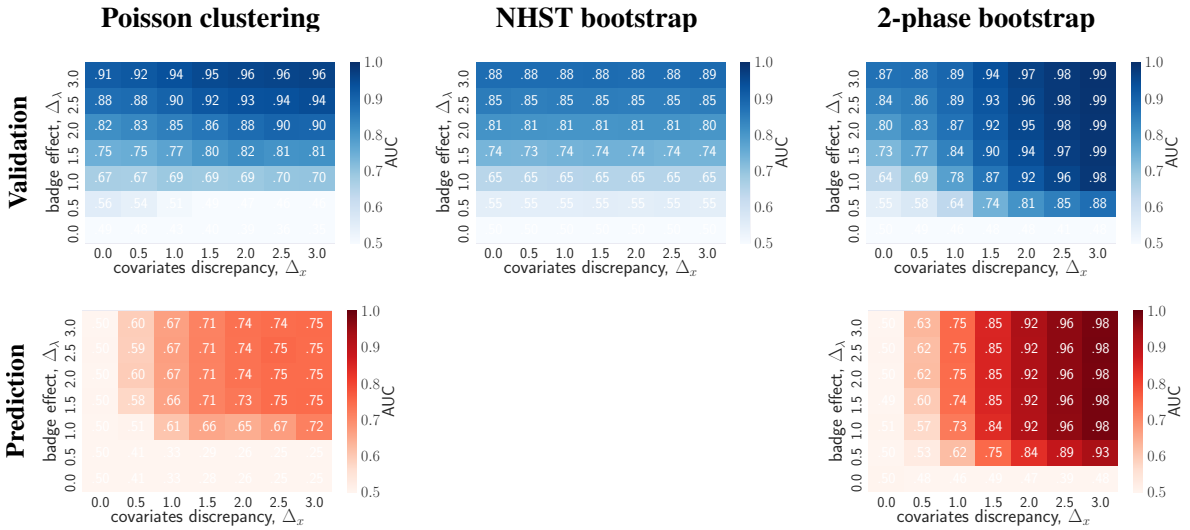


Figure 2: Performance (average AUC) of our methods on synthetic data against badge effect (Δ_λ) and covariates strength (Δ_x). The top row shows the validation of badges' causal effect on users with badge (*i.e.*, having sufficient $\{t_u\}$). The bottom row shows the performance for new users (*i.e.*, with only $\vec{x_u}$ employed).

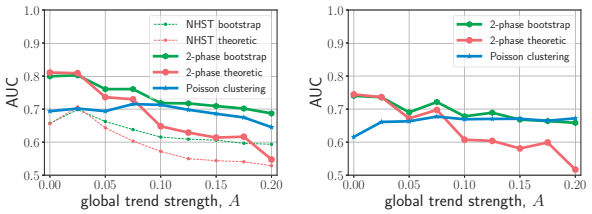


Figure 3: Robustness against temporal fluctuations (*e.g.*, global linear trend; $\Delta_\lambda = \Delta_x = 1, \pi = 0.5$).

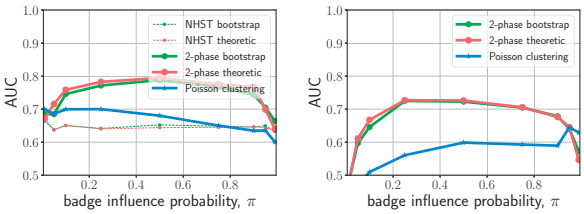


Figure 4: Sensitivity to class imbalance ($\Delta_\lambda = \Delta_x = 1$).

used the following user features and statistics (=proxies describing user activeness level):

- user age and location
- total number of visited pages (page views count), posted comments, upvotes and downvotes

User statistics was transformed by applying the following transformation: $x \to \log(x+1)$. From location we extracted city and state names that we independently embedded using a pre-trained word2vec model[6]. Embeddings were clustered separately into 5+5 clusters using k-means and distances to cluster centers were subsequently used as covariates: 5 for city and 5 for state.

User activeness intervals (*i.e.*, time points s_u and e_u) can be extracted from the posting history, *i.e.*, by taking the times of respectively the first and the last post (= question, answer, comment or edit). However, there might be several badges related to the same action type a (for example, *Tag Editor* and *Research Assistant* are awarded for the 1. and 50. wiki tag edits, respectively). In that case, to avoid interactions between effects from different badges, only actions in direct neighborhood of the badge should be considered. In particular, we update s_u and e_u associated with the badge b in a way that the awarding time b'_u of the other badge b' would be located beyond (with a sufficient margin) these limits.

Users who fulfilled the conditions necessary to get the badge before its introduction time τ could not have been influenced by it and therefore, we filter them out. Similarly, we ignore all the users who lost interest in active participation in the community by that time, *i.e.*, whose end time $e_u < \tau$. Finally, users with incomplete records, *e.g.*, missing age or location, were also disregarded.

We demonstrate the effectiveness of the proposed approaches for two sample *threshold badges*[7]:

- *Research Assistant*: awarded to users who edited at least 50 wiki sites describing tags (wiki tag edits). Users with reputation[8] 1500 or higher can perform these actions.

[6] https://code.google.com/archive/p/word2vec/

[7] https://meta.stackexchange.com/questions/67397/

[8] https://stackoverflow.com/help/whats-reputation

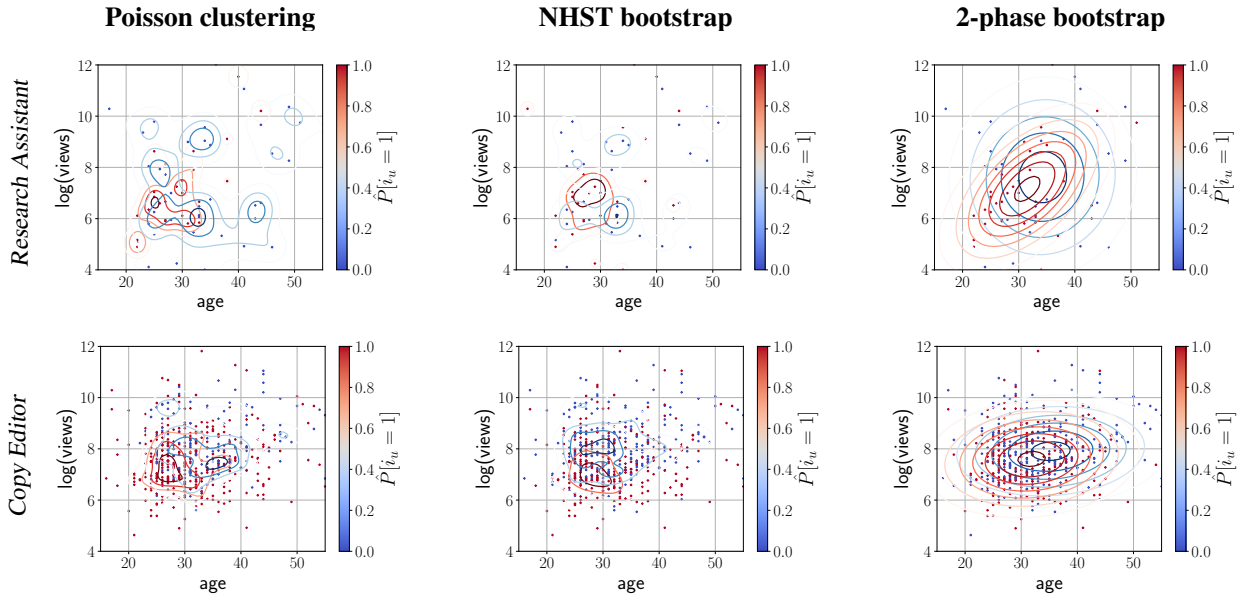


Figure 5: Validation of the effect of two badges: *Research Assistant* (top) and *Copy Editor* (bottom). Users are projected onto a two-dimensional space of *age* and *log number of views*, with badge effect represented with color (`red`=influenced, `blue`=badge awarded by chance).

– *Copy Editor*: awarded to users who performed a total of 500 post (*e.g.*, question or answer) edits. Users with reputation 100 or higher can perform these actions.

Results. Figure 5 illustrates validation results from Poisson processes clustering and 2-phase bootstrap alongside with intermediate results from NHST bootstrap. The classification results from different methods agree (=badge effect probability either larger than 0.5 or smaller than 0.5 in both cases) to a high degree. For example, for *Research Assistant* we observe 70% agreement between Poisson clustering and 2-phase bootstrap. With p-value < 0.001 we can reject the hypothesis that it happens by chance. Similarly, for *Copy Editor* we report p-value $= 0.016$.

The validation results suggest that only about half (*i.e.*, 58% for *Research Assistant* and 47% for *Copy Editor* according to 2-phase bootstrap) of the users intentionally performed actions needed to receive the badge. Prediction results are less conclusive. However, in 2-phase bootstrap classification for *Research Assistant* and for *Copy Editor* we got respectively 53% and 39% of users potentially attracted to the badge (only users with sufficient reputation included), Poisson processes clustering classified all new users as unlikely interested in the badges. We presume that this can happen due to differences between data distributions of users with and without badge that are handled differently by the methods. In particular, it can be a problem if users without badge are predominant in the used data set.

Examination of fitted models can give deeper insights into how user characteristics relate to badges influence. In particular, we ranked covariates according to *Kullback-Leibler divergence* between both user classes ($i_u = 0/1$) and reported means of the respective distributions. We observed that features derived from location best discriminate between classes. For example, we discovered that users located in the USA were getting badges more often by chance – due to their natural high activeness. On the other hand, users from East Europe and India were more mercenary – their behavior was more often driven by the perspective of a badge reward. Similarly, we found out that younger users on average were more goal-oriented than older ones.

Conclusions

Badges are a popular motivational mechanism used in social media sites. However, due to complexity of these environments, the belief that they really work, *i.e.*, are incentivizing users to perform certain actions, is hard to verify and until recently there were no tools for that. To address this problem we designed and evaluated two approaches to verify individual users attraction towards badges. The proposed methods applied to real data from Stack Overflow let us to gain interesting insights about users who earn badges. In particular, in contradiction to previous beliefs we discovered that many of them receive badges by chance, having no prior intention of it.

Our work can be extended in many ways. For example, it would be interesting to see how more advanced features, like temporal features covering user evolution on early stage, can improve the performance of our methods. Furthermore, we focused our research on threshold badges (that are the most popular ones) but there are many other interesting designs (for example badges associated to limited resources) for which the problem of influence validation remains open. Finally, we believe that our results should affect how badges are designed and help in making them more effective.

Acknowledgements

We thank Manuel Gomez-Rodriguez for inspiring us to perform this research, Eliezer de Souza da Silva for useful discussions, and both of them along with Sean Chester for critical review of the publication.

References

Abramovich, S.; Schunn, C.; and Higashi, R. M. 2013. 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.

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

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

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Bornfeld, B., and Rafaeli, S. 2017. Gamifying with badges: A big data natural experiment on stack exchange. *First Monday* 22(6).

Colquhoun, D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1(3).

Daley, D., and Vere-Jones, D. 2002. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer.

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

Ghosh, A., and McAfee, P. 2011. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World Wide Web*, 137–146.

Gibson, D.; Ostashewski, N.; Flintoff, K.; Grant, S.; and Knight, E. 2015. Digital badges in education. *Education and Information Technologies* 20(2):403–410.

Hamari, J.; Koivisto, J.; and Sarsa, H. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, 3025–3034. IEEE.

Hamari, J. 2017. Do badges increase user activity? a field experiment on the effects of gamification. *Computers in Human Behavior* 71:469 – 478.

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

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

Kusmierczyk, T., and Gomez-Rodriguez, M. 2017. Harnessing natural experiments to quantify the causal effect of badges. *arXiv:1707.08160*.

Lewis, M. 2004. The influence of loyalty programs and short-term promotions on customer retention. *Journal of marketing research* 41(3):281–292.

Liang, D.; Charlin, L.; McInerney, J.; and Blei, D. M. 2016. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, 951–961.

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

Sellke, T.; Bayarri, M. J.; and Berger, J. O. 2001. Calibration of ρ values for testing precise null hypotheses. *The American Statistician* 55(1):62–71.

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

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