# One Size Does Not Fit All: Badge Behavior in Q&A sites

Stav Yanovsky
Ben-Gurion University, Israel
stavyano@post.bgu.ac.il

Nicholas Hoernle
University of Edinburgh, U.K.
n.s.hoernle@sms.ed.ac.uk

Omer Lev
Ben-Gurion University, Israel
omerlev@bgu.ac.il

Kobi Gal
Ben-Gurion University, Israel
University of Edinburgh, U.K.
kobig@bgu.ac.il

## ABSTRACT

Badges are endemic to online interaction sites, from Question and Answer (Q&A) websites to ride sharing, as systems for rewarding participants for their contributions. This paper studies how badge design affects people's contributions and behavior over time. Past work has shown that badges "steer" people's behavior toward substantially increasing the amount of contributions before obtaining the badge, and immediately decreasing their contributions thereafter, returning to their baseline contribution levels. In contrast, we find that the steering effect depends on the type of user, as modeled by the rate and intensity of the user's contributions. We use these measures to distinguish between different groups of user activity, including users who are not affected by the badge system despite being significant contributors to the site. We provide a predictive model of how users change their activity group over the course of their lifetime in the system. We demonstrate our approach empirically in three different Q&A sites on Stack Exchange with hundreds of thousands of users, and we discuss the implications for system designers.

**ACM Reference Format:**
Stav Yanovsky, Nicholas Hoernle, Omer Lev, and Kobi Gal. 2019. One Size Does Not Fit All: Badge Behavior in Q&A sites. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3320435.3320438

## 1 INTRODUCTION

Many online platforms rely on the motivation of volunteers rather than on paid workers to create content [14]. Examples include Wikipedia, Reddit, Question and Answer (Q&A) sites like Stack Overflow, and citizen science platforms in which non-experts collaborate with scientists to accelerate scientific discoveries [20]. Social-media websites also rely, to a large degree, on users for creating content.

Keeping users productive and motivated is essential to the success of such peer production sites [20]. One of the most commonly used incentive mechanisms used by these sites are badge systems, which provide users with credentials that display skills and achievements on the site [4, 18]. Badge systems partition the set of participants into "status classes" that reflect their contributions [13]. When administered successfully, badge systems can influence users' behavior and direct them towards types of activities encouraged by the system designers [3].

Despite the massive use of badges in online communities[1], Q&A sites[2], ridesharing[3] and more, our understanding of the interplay between user behavior and badge design is still lacking.

Much previous work has focused on badges' "steering" effect [2, 17]. That is, users' contribution levels rise as they get closer to the threshold that is required for obtaining the badge, and experience a sharp decline following it, returning to their baseline contribution levels.

In this paper we show that the steering effect is not homogenous, but varies across different types of users. Our data is taken from the *Stack Exchange* platform, which hosts a collection of Q&A websites, each devoted to a different topic and includes hundreds of thousands of users. We focused on the largest project of the platform, the programming-related *Stack Overflow* (we also analysed data from several of the smaller projects, like Ask Ubuntu and TeX-LaTeX but as results are similar we will often refer only to Stack Overflow).

We examine a common and general task on Stack Overflow: editing others' posts for corrections and clarifications. We show that the user population can be clustered into 3 separate groups, differentiated by the frequency and intensity of their work on the task. We show that users in each of these different groups experience steering in a different way, and we examine their short and long term reaction to receiving badges. In particular, we find that some users do not return to their baseline levels of contributions after receiving the badge. For these users, badges act as a *catalyst* for long-term activity on the platform, creating a sustained level of activity over an extended period of time.

We provide a computational model for predicting whether a user will decrease her level of contribution to a lower activity group on the site following a badge award. This model can inform the design of personalised intervention methods to increase the contributions of such users. Our work has insights for system designers in showing that badges are not a "one size fits all" incentive and it suggests ways to adapt existing badge designs to the diversity of user behavior.

[1]http://duolingo.wikia.com/wiki/Achievements
[2]https://askubuntu.com/help/badges
[3]https://blog.lyft.com/badge-glossary/


*UMAP '19, June 9–12, 2019, Larnaca, Cyprus*

ACM ISBN 978-1-4503-6021-0/19/06…$15.00
https://doi.org/10.1145/3320435.3320438

## 2 RELATED WORK

Badge design and the effects of badges on people's interactions with online systems has been studied in the social and computational sciences. Hickey et al. [12] outlined key guidelines for successful badge design, such as transparency (the badge system should be known and understood by all users, badges should be visible), interactions (badge systems are more successful in settings where there is a high degree of interaction between participants), and uniqueness (badge systems should be the sole incentive mechanism in the domain setting).

A few studies model the influence of badges on user behavior in social media and Q&A sites [2, 4, 17, 21]. Most central to our work are the studies by Anderson et al. [2] and Li et al. [17] which describe the "steering" phenomena towards a badge boundary: As users approach the threshold of the required number of actions needed to earn a badge (day zero), they increase their contributions needed for the badge.

Anderson et al. [2] present a mathematical model which describes the deviation of distribution over user actions before and after receiving the badge. They use the model to demonstrate the steering effect of badges on user voting behavior on Stack Overflow. First, we show that the steering effect is not homogeneous, it differs across different types of users. Second, we study the long-term effect of badges over the lifetime of interaction of users in the system.

Several works have studied badges in the context of academic courses. Anderson et al. [3] studied badge design and its effect on student behavior in a large student forum in a massive open online course (MOOC). They showed that placing several badges of smaller value that are well dispersed in the course can be more effective than having a single badge of higher value.

Hakulinen et al. [10] showed that rewarding students taking a computer science course with achievement badges motivated students and encourage desired study practices. Charleer et al. [5] studied different visualizations of badges that reward students' forum activity in a course. They compared personal dashboards, where students can observe each other's badge achievements, and an augmented version in which students could discuss the badge achievements with each other. They showed that the personal dashboard improved awareness of the course's goals, while the interactive visualization improved the students' collaboration and reflection on the coursework.

Abramovich et al. [1] used an intelligent tutoring system that notified students whenever they earned a badge and explained the reason for earning it. This approach led to an improvement in students' engagement and a decrease in counter-productive behavior, when compared to badge-less tutoring systems.

Badges have been used in gamified apps, as systems designers use game design elements to improve user engagement and experience [8, 11, 18]. Common gamification elements includes the use of points, levels, leaderboards, time constraints, badges and more [18]. Jia et al. [15] present a survey study investigating the relationships among individuals' personality traits and perceived preferences for various gamification elements.

Badge design has also been studied from a game theoretic perspective [9]. Immorlica et al. [13] studied badge design mechanisms

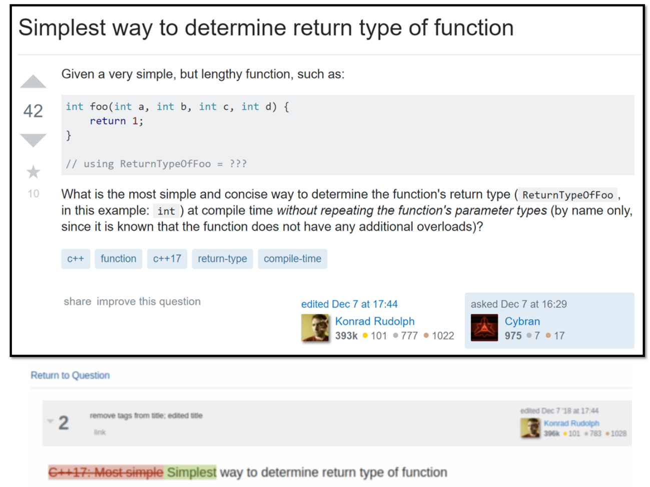


**Figure 1: An example of a post in the Stack Overflow project (top) and an edit activity to the post (bottom)**

aiming to maximize the total contributions made to a website. Users exert an effort (which carries a cost) to contribute and, in return, are rewarded with badges. Badge valuations are determined by the number of users who earn each badge. **?** ] considered the role of badge design as what feedback to provide to two agents in a two-round contest. The agents can expend some amount of effort in each round, with a noisy mapping between effort and score. Both papers characterize the equilibrium strategies that need to hold in their respective model.

Finally, several works criticized the use of badges as an incentive mechanism [7]. In particular, Kobren et al. [16] found that students tend to drop out of e-learning systems just after obtaining the necessary amount of questions to achieve the relevant badge.

## 3 EMPIRICAL METHODOLOGY

Our empirical methodology analyzes data from online platforms that deploy badge incentive schemes and aims to model and understand how different groups of users react to the badge design.

### 3.1 The Stack Exchange Platform

Stack Exchange (SE) is a network of 173 question-and-answer (Q&A) websites on topics in diverse fields, in which users post and respond to questions[4].

For the investigation that follows, we chose the following three projects in SE which vary widely in their topics and in the number of active users.

(1) Stack Overflow (about 9,000,000 users) – deals exclusively with programming and is the biggest and most popular site on SE;
(2) Ask Ubuntu (about 474,000 users) – a site for users and developers of the Ubuntu operating system;
(3) TeX-LaTeX (about 67,000 users)– a site for users of TeX, LaTeX, ConTeXt, and related typesetting systems.

[4]https://stackexchange.com/. User data from SE is freely available.

The primary purpose of each SE site is to enable users to post questions so other users can answer them. Users can also edit and comment on each other's posts. This allows users to correct existing posts or to provide insights about the post content. Users can also vote for posts, providing a reputation mechanism. Users can unlock privileges on the site by increasing their reputation, which can be done by performing actions such as posting questions or answers, voting on posts and editing existing posts. Figure 1 shows an example of a post (top) and edit action (bottom) in SO. Lastly there are a multitude of casual users who access content on the website but do not actively contribute to it.

All SE projects employ badges to incentivize contributions by users. There are more than 100 different badge types in SE and they can broadly be classified as rewarding either the quantity or quality of users' contributions. An example of the latter are "Announcer Badges" that reward users for posts that are visited many times by other users. An example of the former are "Edit Badges" that reward users for making corrections and comments to existing posts. All SE sites divide each badge type to three values in increasing order of importance: bronze, silver and gold. Moreover, SE has aliases for the different type of badges that one can earn. For example, edit-type badges in SE sites use the aliases "Editor" for a bronze badge value, "Strunk & White" for a silver badge value, and "Copy Editor" for a gold badge value.

## 3.2 Research Questions

Our research focuses on the way badges affect different users. We are particularly interested in the *steering* phenomenon identified by Anderson et al. [2], where users tend to increase their contribution rates as they approach the badge goal. We are also interested in modeling how badges affect people's long term behavior on the platform.

We study the following research questions:

(1) Does the steering effect identified by Anderson et al. [2] extend to other types of actions and SE projects, beyond voting actions in Stack Overflow (SO)?
(2) Do different user populations experience steering in different ways?
(3) How do badges affect the long term behavior of individual users throughout the lifetime of their interaction in the system?

We focus on analyzing badge behavior for edit actions because they represent a quantitative family of badges. Frequent edit type actions in SE include correcting grammatical errors or misspellings in a post, or adding explanations to the existing post content. The thresholds for achieving the bronze, silver and gold badges are a single edit action, 80 edit actions, and 500 edit actions respectively. These threshold values are standardized across all of the SE projects. Our hypothesis was that different types of users "steer" differently, i.e., they vary in the extent to which they respond to badges. Moreover, we believe that individual users vary in how the badges affect their behavior throughout the course of their lifetime on the system.

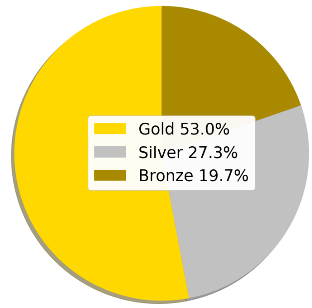


**Figure 2: Percentage of edit-actions contributions for winners of bronze, silver and gold badges**

# 4 ANALYSIS AND RESULTS

Our analysis is based on data of user interactions on the SO project from September 2008 and up to August 2018, including about 8,500,000 edit actions. Figure 2 shows the percentage of contributions made by winners of the different badge types. We focus the analysis of this paper on the 14,276 users who achieved the silver badge; 2,687 of these users went on to achieve the gold badge. Together, this group made the vast majority of edit contributions to the site. Understanding how these users behave can inform the design of future incentive mechanisms in the site. Although winners of the bronze badge make up the majority of the user population (), they do not provide a substantial contribution of edit actions to the site. In the Discussion & Future Work Section we explore how the badge system might be designed to encourage more participation from the users who only achieved the bronze badge.

## 4.1 Question 1: Does Steering Generalize?

Anderson et al. [2] have demonstrated the steering effect on the SO site. In this section we study whether steering generalizes to Edit badges and to other SE projects. Figure 3 shows the average number of edit contributions as a function of the number of days from day-zero (the day in which the silver badge was obtained by the user) for all three projects: Stack-Overflow (top), Ask Ubuntu (middle) and TeX-LaTeX (bottom) SE sites. The negative numbers to the left of day-zero give the time in days prior to obtaining the badge. Accordingly, the region to the right of the axis, with positive numbers, is the time after getting the badge.

As shown by the figure, in all projects, users exhibit a sharp rise in activity as they approach day zero, and following this day they exhibit a steep decline in contribution, returning to their default rates of activity. The figure confirms that the steering effect identified by Anderson et al. [2] holds in other projects as well. Moreover, the steering effect was not limited to only voting actions and badges but appears to hold for editing action and badges as well. The question that flows directly from this result is whether steering is a "one size fits all" phenomena, that is, do all users exhibit the same kind of steering behavior?

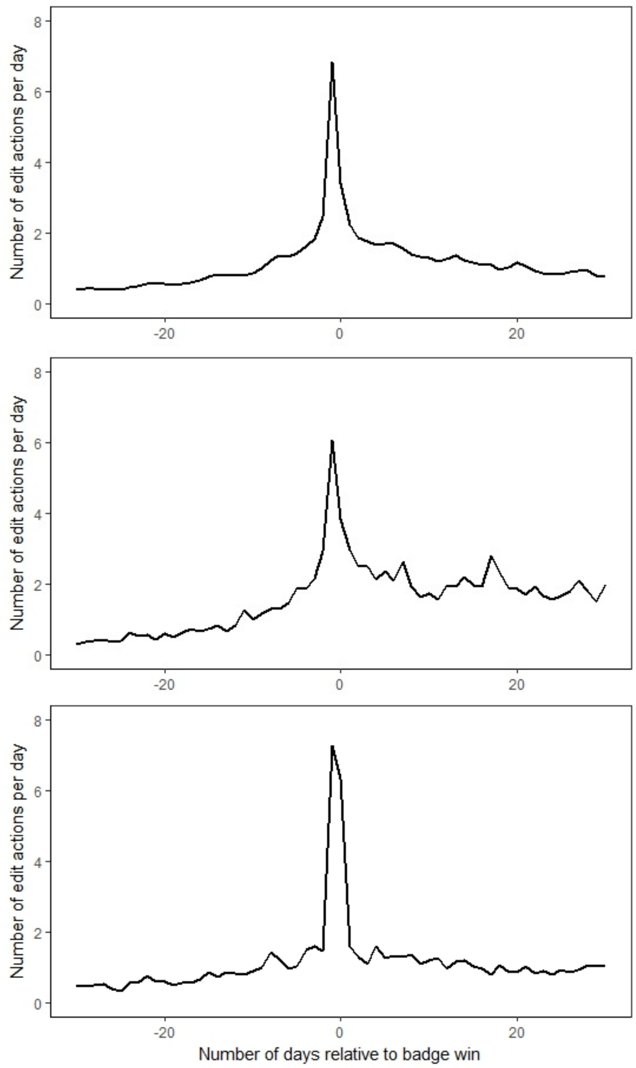


**Figure 3: Average number of edit contributions as a function of distance from day-zero for obtaining silver badge on the Stack-Overflow (top), AskUbuntu (middle) and TEX-LATEX SE sites (bottom).**

## 4.2 Question 2: Is Steering One-Size-Fits-All?

In this section we study whether steering effects differ between different types of users, as exhibited by their behavior on the site. Intuitively, users with similar numbers of contributions may still exhibit widely different activity styles. For example, consider two users; one of them performs 5 edit actions each day of the week and the other performs 35 edit actions on Sunday nights. In total, both users contribute an equal number of edit actions per week, but clearly, they exhibit different behavior patterns on the site.

To distinguish between such users, we define two measures: The amount of edit contributions they make in a given time period, and the frequency with which they make contributions in that time period. We chose these two measures because (1) they provide a general description of user activity in SE that does not depend on the action type itself (e.g., the number of characters changed or added in an edit activity); (2) they provide a simple and succinct way to differentiate between user behavior in the site.

4.2.1 ***Measuring User Activity***. We used the following measures to describe user activity:

**Work Consistency:** The median number of days spent editing in a week.

**Work Intensity:** The median number of edits that a user makes in a day, given the user makes at least one edit.

For example, a user who was active for three days in the first week of activity, five days in the second week of activity, and three days in the third and final week of activity will have a consistency value of 3. Similarly, a user who produced two edit actions in the first day of activity, ten edits in the second day of activity, and five edits days in the third and final day of activity will have an intensity value of 5. We considered the median number of edit actions rather than the mean because the distribution over edit actions per day is right-skewed and is highly affected by outliers (e.g., consider a user who completes 2 edits per day for 25 days but has a single high interaction day of 30 edits. This user meets the 80 edits required to achieve the Silver badge but her mean edits per day is 2/3 larger than the median).

4.2.2 ***Inferring User Groups***. Using the notions of work consistency and intensity, we wish to group users into distinct clusters of activity. In order to do so, we utilize the k-means algorithm. Figure 4 plots the work consistency and intensity for the gold and silver users in SE (recall this group contributes the vast majority of edits in the system). The algorithm used the two measures to cluster users into three groups of activity: low, medium and high. Groups are distinguished in the figure using colors and boundary curves.

The default distance metric used in k-means is euclidean distance between datapoints and cluster centers. However this does not constitute a good metric for the purposes of differentiating between groups with different levels of contributions. For example, when using euclidean distance, if we consider a cluster with (consistency, intensity) centroid $(4, 5)$, and two users with (consistency, intensity) measures of $(1, 5)$ and $(7, 5)$, the users would exhibit the same distance from this cluster center. However they display different activity levels. The first user works one day a week and would complete an expected total of 5 edits per week, while the second works every day in the week and would complete an expected total of 35 edits per week.

To this end, we define a custom distance that captures the expected number of posts per week directly. For two users $a$ and $b$ and their respective intensity and consistency ($I_a, C_a, I_b$ and $C_b$), the distance $d$ is defined as:

$$d(a, b) = ABS(I_a \times C_a - I_b \times C_b) \qquad (1)$$

The group centers can thus also be interpreted in terms of expected number of edits per week.

The clusters are formed in a transformed parameter space using the following steps. First, we drop the users with consistency and intensity scores greater than the 99.9 percentile in each case. This corresponds to 15 users who all had an intensity greater than 20.

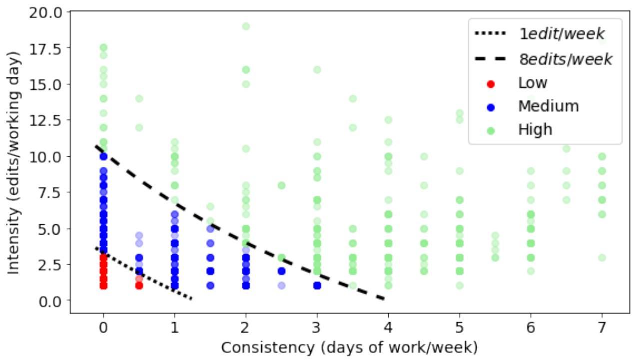


**Figure 4: Scatter plot of user activity showing three user groups revealed by k-means (K=3). Groups are distinguished using colors and boundary curves.**

| | Low | Medium | High |
|---|---|---|---|
| **Silver users** | 10,022 | 1,380 | 287 |
| **Gold users** | 1,198 | 1,119 | 370 |
| **Total** | 11,220 | 2,499 | 657 |

**Table 1: Number of gold and silver users in each activity group**

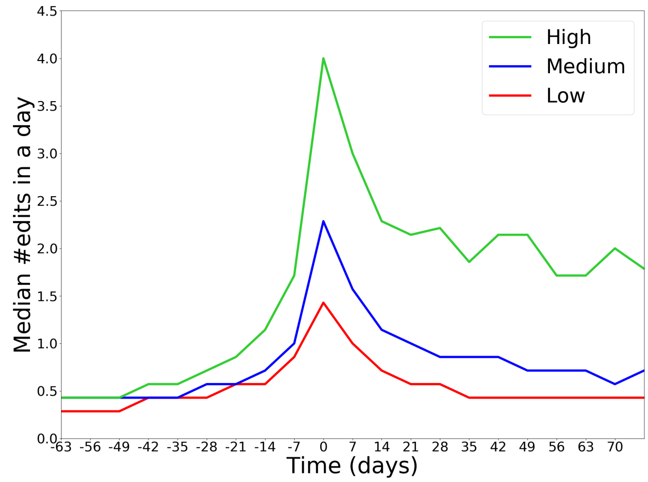


**Figure 5: Median number of edits per day, centered around the day-zero for achieving the silver badge**

Second, we normalize the data by dividing by the maximum value and adding 1 to offset the effect of 0 values.

We choose $k = 3$ to facilitate the interpretation of the clusters. We aim to describe general trends in the data while still accounting for the fact that users are interacting with the system in unique ways. Increasing the cluster parameter $k$ to 4 simply had the effect of splitting the high-activity group into two, thus complicating the further analysis unnecessarily.

Using the modified distance metric, the k-means algorithm reveals the following three types of user groups. The low activity group describes "dabbler" users whose activity is characeterized by low consistency and intensity levels (contribute on average less than a single edit per week). The medium activity group describes users who exhibited a medium level of intensity and consistency (contribute on average between one and 8 edits per week and rarely work more than 4 days for any given week). The high activity group describes "busy bee" users who exhibited a high level of intensity (contribute on average more than 8 edits per week and regularly work on more than 3 days in any given week). Returning to our example from above with the two users described by (consistency, intensity), we can see that the first user is in the medium activity group and the second user is in the "busy bee" group.

Table 1 shows the number of silver and gold users in each activity group. As shown in the table, low-activity users make up the vast majority of the user population, followed by the medium and high activity user groups. Gold users make up just 16% of low activity users, 44% of the medium user group, and over a half of the high user group. Thus, despite the lower rate of contributions exhibited by the low/medium activity groups, they still make up a substantial part of the contribution.

#### 4.2.3 *Separating the Badge Effect.*

Figure 5 plots the contributions of the different engagement groups over time, relative to day zero, when the silver badge was awarded to the user. As shown in the figure, several days before day zero, there is no discernible difference in the contribution levels of the three groups, as their usage pattern seems to be identical across the clusters.

However, in the few days just before day zero, high-activity users experience a sharp rise in the number of edits per day leading into the silver badge and maintain this high rate of editing for a number of weeks (!) after the receiving the badge. The behavior of these users runs counter to Anderson et al. [2]'s prediction that after receiving a badge, users will return to their default levels of activity. In contrast, the other groups are far less affected by the badge design. Both low and medium-activity groups exhibited a smaller jump in contributions prior to day-zero (with low-activity also smaller than medium-activity), and a steeper decline after this day. However, medium-activity users did settle on contribution levels slightly above their previous, pre-badge, default level, while low-activity users returned to their previous work habits. We recall that the low activity group is the largest and it therefore dominates the trends when all of these users are aggregated. It is only when we analyze these groups individually that this nuanced behavior becomes apparent.

At the peak of the contribution level, there is a highly statistically significant difference between the three levels of contributions; $p \ll 1 \times 10^{-4}$ one-way ANOVA. Before obtaining the silver badge the difference between the levels of contributions is not statistically significant (30 days before day zero - $p = 0.206$ one-way ANOVA), while after obtaining it, the difference stays significant (30 days after day zero - $p \ll 1 \times 10^{-4}$ one-way ANOVA).

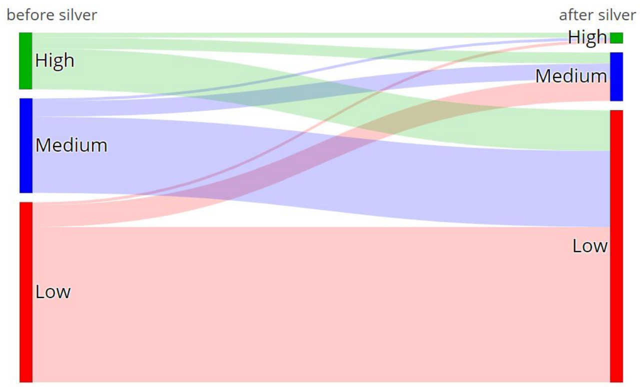


**Figure 6: Flow between user groups before (left) and after (right) getting the silver badge**

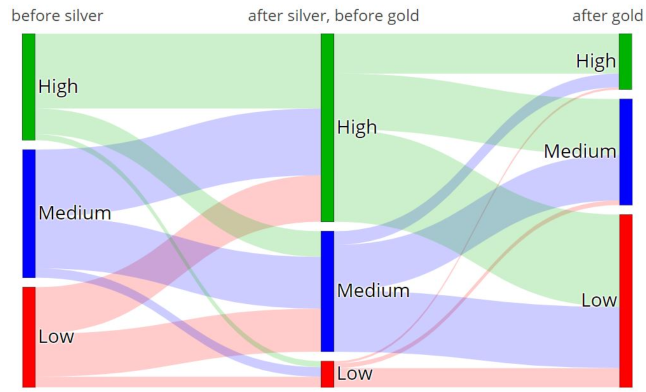


**Figure 7: Flow between user groups before getting the silver badge (left), after getting the silver and before getting the gold badge (middle) and after getting the gold badge (right).**

### 4.3 Question 3: How do Long Term Group Dynamics Change in the Presence of Badges?

In this section we address our third research question, namely how users change their behavior over time. To this end we track whether and how low, medium, and high-activity users change group types before and after receiving badges in the system. Thus, we divide users into groups for different parts of their life cycle in the system (before badge/after badge, etc.).

We begin by tracking the long term behavior of users who received a silver badge but not the gold badge (these users are responsible for 27% of user contribution on the site, see Figure 2). Figure 6 is a Sankey diagram tracking the flow of these users between the different group types. As can be seen, the vast majority of users (including medium and high activity users) became, once the badge was awarded, low activity users. Only a tiny minority of low and medium users became high-activity users. The behavior of these users agrees with the theory of steering, in that they returned to their usual work patterns following the silver badge acquisition. In the discussion section we discuss the implications of this behavior to facilitating badge design.

We now turn to track the long term behavior of users who received the gold badge (these users are responsible for 53% of user contribution on the site). Figure 7 shows the flow between groups for these users before obtaining the silver badge (left), after obtaining the silver and before obtaining the gold badge (middle) and after getting the gold badge (right). As shown by the figure, the user shifting between groups is quite different than that of users who did not obtain the gold badge (Figure 6). The users depicted here did not return to their normal routine once they achieved a silver badge. Instead, they generally increased their activity – an overwhelming majority either stayed at the same activity level or increased it, and for the medium and low-activity users, a sizable majority strictly increased their activity levels to become high-activity users. However, once the gold badge was achieved, most of the users reverted to the same steering behavior identified by Anderson et al. [2], and their activity levels decreased significantly.

## 5 PREDICTING USERS' BEHAVIOR

We turn to the task of predicting the future behavior of a user given the activity in the system prior to receiving the silver badge. We wish to understand which users are susceptible to decreasing their work habits after obtaining the badge, so that we can target these users and tailor incentive solutions for them in order to increase their motivation (see Discussion Section). Therefore, it is critical to know in advance which users will decrease their work on the platform and which users will be maintaining the same levels of work.

### 5.1 Feature Extraction for Prediction Task

A user is represented by a vector that includes three distinct families of predictor variables: **user features**, **edit features** and **temporal features**.

- **User Features** include features specific to the user such as her age, length of activity history in the system as well as the number of other SO badge achievements won prior to the date when the silver badge was awarded.
- **Edit Features** included features that summarize the user's edit history. These features include the ratio of edit actions to other actions performed on the system, statistics about what part of the posts they edit (e.g. title or content) and how long are the comments describing each edit.
- **Temporal Features** included features summarizing the user's consistency and intensity measures from different periods of time from the user's interaction history. We represent the history as a vector of mean consistency values for each week of the user's lifetime in the system, up to day zero (and similarly for intensity values). To measure changes in these two metrics, we average the consistency and intensity through time for 3, 5 and 10 weeks prior to achieving the badge and the 3, 5, and 10 weeks of a user's activity in the system . For example, for a given consistency history $(c_1, \ldots, c_n)$ of $n$ weeks of activity, these features average the consistency values $(c_1, \ldots c_3)$ for the first 3 weeks of activity, and similarly for the first 5 and 10 weeks of activity prior to

| Features | # Features | Accuracy | F1 Weighted | ROC AUC | Confusion Matrix |
|---|---|---|---|---|---|
| **User** | 10 | 0.704 | 0.693 | 0.640 | [7987 1549]<br>[2700 2140] |
| **Edits** | 9 | 0.680 | 0.633 | 0.569 | [8680 **856**]<br>[3739 1101] |
| **Temporal** | 40 | 0.835 | 0.839 | **0.875** | [7174 2362]<br>[**12** 4828] |
| **User + Edits + Temporal** | 59 | **0.837** | **0.842** | **0.875** | [7236 2300]<br>[42 4798] |

**Table 2: Prediction results for user decreasing contribution levels after day zero, using different combinations of features.**

the badge. We also create features for 3, 5 and 10 weeks of the user's interaction history prior to receiving the badge. These features average the consistency values $(c_{n-2}, \ldots, c_n)$ for the last 3 weeks of activity before day zero, and similarly for the last 5 and 10 weeks of activity. Similar features are defined for history relating to intensity users. This allows the prediction to harness relative changes in the user's behavior at different points in time relative to day zero. Another important feature in this family of features is the user's activity group prior for obtaining the badge.

The prediction task is whether the user will decrease her contributions and move to a lower group type after receiving the silver badge. Specifically, we predict whether high activity users descend to the medium activity group, and whether medium activity users descend to the low activity group.

## 5.2 Prediction Results

We used the XGBoost classifier algorithm [6] for this prediction task, and tried different combination of the following parameters: the number of used trees, the maximum depth of the trees and the learning rate of the algorithm. We used ten fold cross validation, with standard deviation of results between runs smaller than 0.01 for all measures. As can be seen in Table 2, in most regards, using all feature types produced the best results. However, note that most of the prediction quality comes from using the temporal based features. The user and edit features have relatively weak prediction ability (a combination of them showed a negligible increase in prediction ability), and using the temporal variables alone seems to give excellent results without requiring any extensive knowledge of the users themselves or their particular editing habits.

Using the user and edit features alone led to a rather small number of errors in one direction – fewer people were mistakenly predicted to decrease their activity, when they did not (false positive). However, using the temporal features alone, while increasing false positives, almost eliminated the error of predicting people will not decrease their activity when the did (false negative). When trying to prevent people from decreasing their activity, false negatives are more important to focus on, because presumably, many engaged users will brush off attempts to engage them further, while users who are not targeted to prevent their dropping-out, are forever lost to the system. This provides further support for the predictive model.

# 6 DISCUSSION & FUTURE WORK

Our main results in this paper elaborate and expand previous research on badges (and in particular, Anderson et al. [2]):

(1) Reaction to badges varies greatly between different user populations. In particular, large sections of users (e.g., our low-activity ones) register a very small reaction to badges at all, while others show a reaction that is at odds with model predictions, as they *increase their work after receiving the badge.*
(2) Engagement patterns of users can be an effective predictor, at least to some degree, of future badge reception (e.g., high-activity users and the gold badge). Classification of users based on their working habits may be beneficial for understanding which people might benefit from different incentives.

We observe a far more complex interaction between badges and user behavior than noted previously. Our results indicate that while for many users, working intensely to receive a badge can be a one-time thing, for some users (who are the most productive ones, from the platform's point of view), badges have a different meaning. These users' behavior seems to indicate that once they receive their first meaningful badge, it encourages them to participate in the badge environment, and they want to achieve more badges, until there are no more to achieve. The badge seems to be the *catalyst* for such a process, as prior to being awarded the silver badge, these users had much lower activity levels. However, even for these types of users, the badge system is meaningful, as once they have received a gold badge, they slowly drop off the system.

Our observations here could be applied in different ways for different use-cases. When a high rate of participation is needed, it is clear that badges are failing to engage vast numbers of users, in particular the low-activity users, and more importantly – those that do not even reach the stage of silver badges. Perhaps a different set of incentives might be needed for these users. On the other hand, badges are much more effective in motivating persistent behavior from a subset of users, and there is potential for badge behavior to focus on these users. For example, it may be beneficial to lower the threshold for awarding badges, to allow for this activity catalyst to reach users who may be "sparked" by it, but prior to receiving the badge, had such a low activity profile that they did not even reach the silver badge threshold. This would allow for an earlier identification of the other engagement groups, and hence to allow for focusing on the different incentives needed for each of these populations. Similarly, medium-activity users are increasing their activity after they receive the silver badge, taking quite a while to return to usual work patterns. Perhaps if the next badge was not so distant (500 edits for gold vs. only 80 for silver), they might have seen the badge goal as reachable, and become high-activity users, working to achieve it.

The intricate interaction we uncover between badges and user behavior calls for much further research, and there is plenty left to do. For example, the role of multiple badges is not yet fully understood. We are extending our work to other types of badges (in particular, qualitative ones), and examine personalized badge structure, which could hopefully be tested in real-world settings. We intend to experiment with different badge design schemes for engaging student learning in a soon-to-be-released MOOC. Our long term goal is to direct system designers on how to design a badge system optimally for a given platform. Also, we are studying how to design intervention mechanisms that target individual users who are predicted to decrease their contribution level. To this end it is necessary to reason about the trade-off that is made between interrupting or frustrating an engaged user and between intervening with a user who might be disengaging with the platform [19].

## 7 ACKNOWLEDGMENTS

Thanks very much to Stack Overflow for making available the data on which this research was based. Nicholas Hoernle is supported by a commonwealth scholarship. Stav Yanovsky is supported by a grant from the Israeli Science Foundation number 773/16.

## REFERENCES


[1] 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.

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 95–106.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*. ACM, 687–698.

[4] Huseyin Cavusoglu, Zhuolun Li, and Ke-Wei Huang. 2015. Can Gamification Motivate Voluntary Contributions?: The Case of StackOverflow Q&A Community. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (CSCW'15 Companion)*. ACM, New York, NY, USA, 171–174. https://doi.org/10.1145/2685553.2698999

[5] Sven Charleer, Joris Klerkx, Santos Odriozola, Jose Luis, and Erik Duval. 2013. Improving awareness and reflection through collaborative, interactive visualizations of badges. In *ARTEL13: Proceedings of the 3rd Workshop on Awareness and Reflection in Technology-Enhanced Learning*, Vol. 1103. CEUR-WS, 69–81.

[6] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[7] Edward L Deci, Richard Koestner, and Richard M Ryan. 2001. Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of educational research* 71, 1 (2001), 1–27.

[8] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. 2011. Gamification. using game-design elements in non-gaming contexts. In *CHI'11 extended abstracts on human factors in computing systems*. ACM, 2425–2428.

[9] David Easley and Arpita Ghosh. 2016. Incentives, gamification, and game theory: an economic approach to badge design. *ACM Transactions on Economics and Computation (TEAC)* 4, 3 (2016), 16.

[10] Lasse Hakulinen, Tapio Auvinen, and Ari Korhonen. 2015. The Effect of Achievement Badges on Students' Behavior: An Empirical Study in a University-Level Computer Science Course. *International Journal of Emerging Technologies in Learning* 10, 1 (2015).

[11] Juho Hamari, Kai Huotari, and Juha Tolvanen. 2015. Gamification and economics. *The gameful world: Approaches, issues, applications* 139 (2015).

[12] Daniel T Hickey, J Willis, and J Quick. 2015. Where badges work better. *ELI. Retrieved January* 31 (2015), 2017.

[13] 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 2015, Florence, Italy, May 18-22, 2015*. 473–483.

[14] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on World wide web*. ACM, 143–154.

[15] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-targeted gamification: a survey study on personality traits and motivational affordances. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2001–2013.

[16] Ari Kobren, Chun How Tan, Panagiotis Ipeirotis, and Evgeniy Gabrilovich. 2015. Getting more for less: Optimized crowdsourcing with dynamic tasks and goals. In *Proceedings of the 24th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 592–602.

[17] Zhuolun Li, Ke-Wei Huang, and Huseyin Cavusoglu. 2012. Quantifying the impact of badges on user engagement in online Q&A communities. (2012).

[18] Katie Seaborn and Deborah I Fels. 2015. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies* 74 (2015), 14–31.

[19] Avi Segal, Kobi Gal, Ece Kamar, Eric Horvitz, and Grant Miller. 2018. Optimizing Interventions via Offline Policy Evaluation: Studies in Citizen Science. (2018).

[20] Robert Simpson, Kevin R Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*. ACM, 1049–1054.

[21] Jiawei Zhang, Xiangnan Kong, and Philip S Yu. 2016. Badge system analysis and design. *arXiv preprint arXiv:1607.00537* (2016).