Received: 20 December 2019 | Revised: 5 August 2020 | Accepted: 8 August 2020

DOI: 10.1002/asi.24409

RESEARCH ARTICLE

JASIST WILEY

One size does not fit all: A study of badge behavior in stack overflow

Stav Yanovsky[1] | **Nicholas Hoernle**[2] | **Omer Lev**[1] | **Kobi Gal**[1,2]

[1]Department of Software and Information Systems Engineering, Department of Industrial Engineering and Management, Ben-Gurion University, Beersheba, Israel

[2]School of Informatics, University of Edinburgh, Edinburgh, UK

Correspondence
Kobi Gal, Department of Software and Information Systems Engineering, Department of Industrial Engineering and Management, Ben-Gurion University, Beersheba, Israel.
Email: 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 article 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, for two types of activities (editing and voting on posts).

1 | INTRODUCTION

Many online platforms rely on the motivation of volunteers rather than on paid workers to create content (Ipeirotis & Gabrilovich, 2014). Examples include Wikipedia, Reddit, question and answer (Q&A) sites like Stack Overflow (SO), and citizen science platforms in which nonexperts collaborate with scientists to accelerate scientific discoveries (Simpson, Page, & De Roure, 2014). Moreover, 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 (Simpson et al., 2014). 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 (Cavusoglu, Li, & Huang, 2015; Seaborn & Fels, 2015). Badge systems partition the set of participants into "status classes" that reflect their contributions according to a particular metric (Immorlica, Stoddard, & Syrgkanis, 2015). When administered successfully, badge systems can influence users' behavior and direct them toward types of activities encouraged by the system designers (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014). [The use of badges can also signal expertise or experience to users as well as to wider communities (Hickey, Willis, & Quick, 2015).]

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 (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2013; Li, Huang, & Cavusoglu, 2012). 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 article, we show that the steering effect is not homogeneous, but varies across different types of users. Our data are taken from the *Stack Exchange* (SE) 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 SO (we also analyzed data from several of the smaller projects, like Ask Ubuntu and TeX-LaTeX, but as results are similar we will often refer mainly to SO).

We examine two common and general tasks on SO: voting on other people's posts, and editing others' posts for corrections and clarifications. [We chose these badge types because they represent fundamental and popular activities in SE, hence our insights are representative of the general population of users.] We show that the user population can be clustered into three 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 reactions 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 personalized 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.

This article extends a prior conference submission that focused on the SO project in several ways. First, we extend our results to an additional class of voting actions and show that our model generalizes well to this new class. Second, we examine the interaction between different types of badges. Finally, we extend the related work section to provide more comprehensive background of badge design and user behavior in the literature.

2 | RELATED WORK

Badge design and the effects of badges on people's interactions with online systems have been studied in the social and computational sciences. Hickey et al. (2015) 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). Different sites use different badge designs, each with its own particular purpose (Easley & Ghosh, 2016). For example, some badges award users for contributing valuable content while others award users for being among the most prolific contributors of the site.

A few studies model the influence of badges on user behavior in social media and Q&A sites (Anderson et al., 2013; Cavusoglu et al., 2015; Halavais, Kwon, Havener, & Striker, 2014; Li et al., 2012; Zhang, Kong, & Yu, 2016). Most central to our work are the studies by Anderson et al. (2013) and Li et al. (2012), which describe the "steering" phenomena toward 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. The total amount of actions that are influencing the earning of the badge increases significantly in the days that are prior to earning the badge, in a comparison to earlier days and the days after getting the badge. [Bosu et al., 2013 analyzed the social network inferred from participants' questions and answers on the site, inferring how to obtain reputation scores quickly on the cite. Other works have studied the role of personality traits, as inferred by their SO messages, on their badge behavior (Bazelli, Hindle, & Stroulia, 2013; Papoutoglou, Kapitsaki, & Mittas, 2018). Bornfeld & Rafaeli, 2017 provided a longitudinal analysis of feedback (votes and responses) as a mechanism for increasing contributions among newcomers in five SE communities, including SO.]

Anderson et al. (2013) present a mathematical model that 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 SO, and a few empirical studies followed (Grant & Betts, 2013). However, our results paint a more complicated picture than these, as we show that the steering effect is not homogeneous and differs across different types of users. We also 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. (2014) 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, Auvinen, and Korhonen (2015) showed that rewarding students taking a computer science course with achievement badges motivated students and encouraged desired study practices. Charleer, Klerkx, Odriozola, Luis, and Duval (2013) 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, Schunn, and Higashi (2013) 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 counterproductive 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 (Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011; Hamari, Huotari, & Tolvanen, 2015; Seaborn & Fels, 2015). Gamification is studied from the aspects of psychology (Linehan, Kirman, & Roche, 2015), game theory (Easley & Ghosh, 2016; Jain, Chen, & Parkes, 2009), and economics (Hamari et al., 2015). Common gamification elements include the use of points, levels, leaderboards, time constraints, badges, and more (Seaborn & Fels, 2015). Jia, Xu, Karanam, and Voida (2016) 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 (Easley & Ghosh, 2016). Immorlica et al. (2015) studied badge design mechanisms 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. Aoyagi (2010) 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 (Deci, Koestner, & Ryan, 2001), claiming that badges may undermine intrinsic motivation. In particular, Kobren, Tan, Ipeirotis, and Gabrilovich (2015) 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 | SETTING AND RESEARCH QUESTIONS

The setting for this work is the SE platform, a network of 173 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.

The primary purpose of each SE site is to enable users to post questions so other users can answer them. In order to improve both questions and answers, users can also edit and comment on each other's posts. This allows users to correct existing posts (e.g., in the case of typos) or to provide insights about the post content. Users can also vote (up or down) for posts, providing a reputation score that is displayed on the user profile. By performing actions such as posting questions or answers, voting on posts and editing existing posts, users can increase their reputation score on the site and unlock various privileges (Bosu et al., 2013). Figure 1 shows an example of a post (top) that has been upvoted (left) with an edit action (bottom) in SO. Of course, the multitude of users (as in other crowdsourced projects, such as Wikipedia), are casual ones, who only 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, each divided to three ranks in increasing order of importance: bronze, silver, and gold. To differentiate between them, 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. The badges rewarding voting actions are "Supporter" or "Critic" for one's first vote (depending on if it was an up-voter or

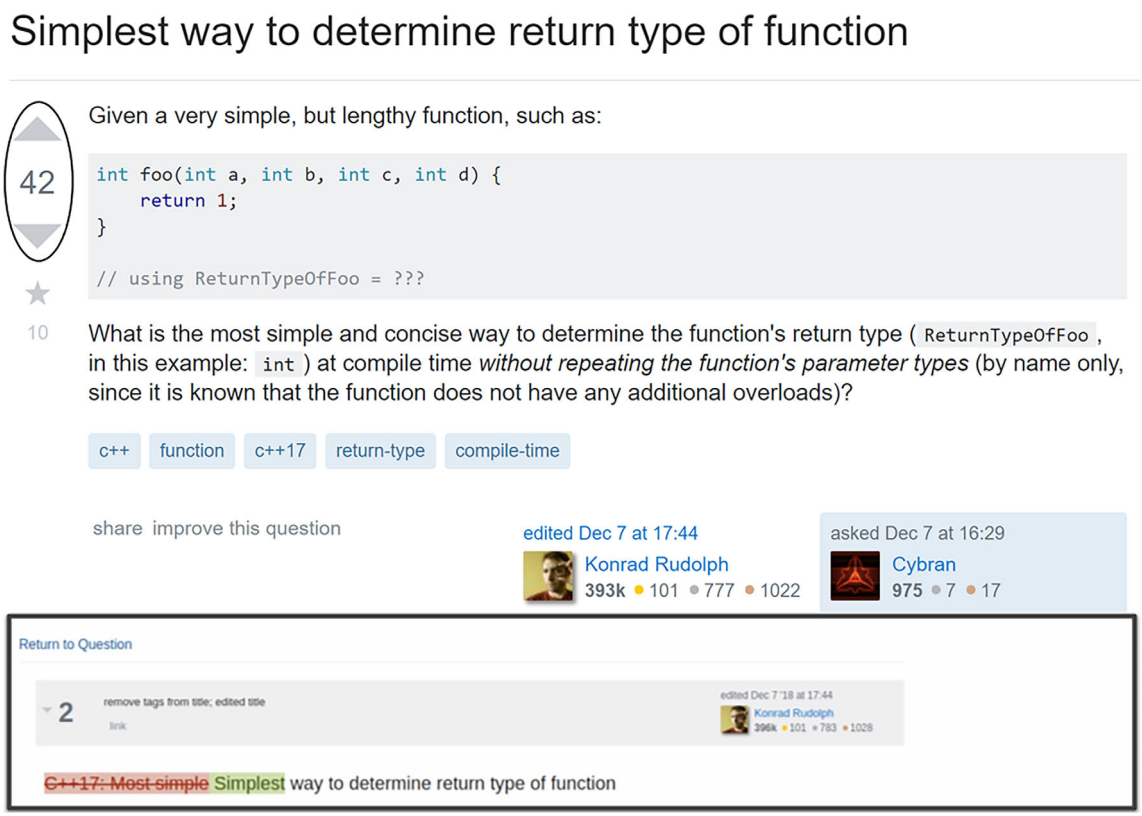

FIGURE 1 An example of a post in the stack overflow project (top), which has been up-voted 42 times (left) and an edit activity to the post (bottom) [Color figure can be viewed at wileyonlinelibrary.com]

down-vote) as the bronze badge; "Civic Duty," given after voting 300 or more times (regardless of the vote type) is the silver badge; and "Electorate," given to users after voting on 600 questions (of which at least 25% are on questions), is the gold badge.

3.1 | The edit badges data

We will initially focus on the edit badge. Our analysis is based on data of user interactions on the SO project from September 2008 and up to May 2019. Figure 2 shows the percentage of edit-action contributions made by winners of the different badge types.

We are particularly interested in 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 vast majority of the user population (more than 90%), they do not provide a substantial contribution of edit actions to the site. In Section 8, we explore how the badge system might be designed to encourage more participation from the users who only achieved the bronze badge.

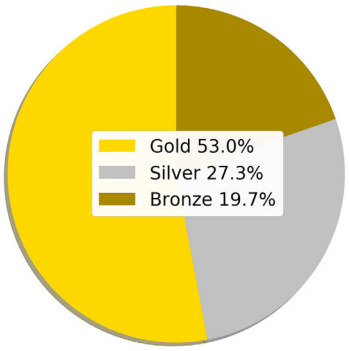

FIGURE 2 Percentage of edit-actions contributions for winners of bronze, silver, and gold badges [Color figure can be viewed at wileyonlinelibrary.com]

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. (2013), 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. Is steering one-size-fits-all? Do different user populations experience steering in different ways?
2. How do badges affect the long-term behavior of individual users throughout the lifetime of their interaction in the system?
3. Does the steering effect identified by Anderson et al. (2013), and our findings on the edit badge, extend to other types of actions and SE projects, beyond voting actions in SO?

We first focus, for the first two questions, on analyzing badge behavior for edit actions. 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, that is, 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. For the third question, we widen our angle to more SE projects and other badges.

4 | QUESTION 1: 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:

1. Work consistency: The median number of days spent editing in a week.
2. Work intensity: The median number of edits that a user makes in a day, given the user makes at least one edit.

We chose these two measures because (a) 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); (b) they provide a simple and succinct way to differentiate between user behavior in the site. For example, a user who was active for 3 days in the first week of activity, 5 days in the second week of activity, and 3 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, 10 edits in the second day of activity, and 5 edits in the third day, and the 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.

4.1 | 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 3 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.

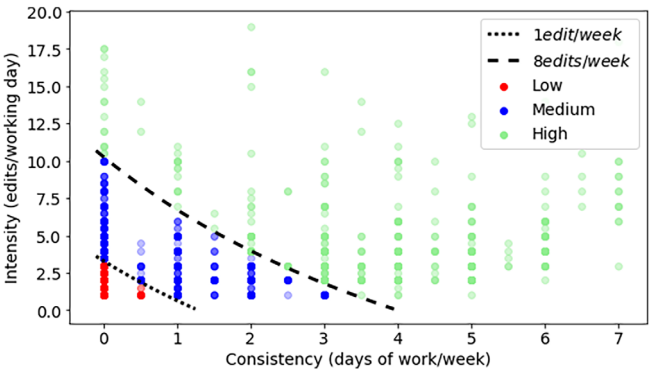

FIGURE 3 Scatter plot of user activity showing three user groups revealed by k-means ($K = 3$). Groups are distinguished using colors and boundary curves [Color figure can be viewed at wileyonlinelibrary.com]

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 1 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:

$$d(a,b) = \mathrm{ABS}(I_a \times C_a - I_b \times C_b). \quad (1)$$

The group centers can thus also be interpreted in terms of common 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. 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 characterized by low consistency and intensity levels (contribute generally less than a single edit per week). The medium-activity group describes users who exhibited a medium level of intensity and consistency (contribute generally 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 generally 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-activity user group, and over a half of the high-activity 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 | Separating the badge effect

Figure 4 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

TABLE 1 Number of gold and silver users in each activity group

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

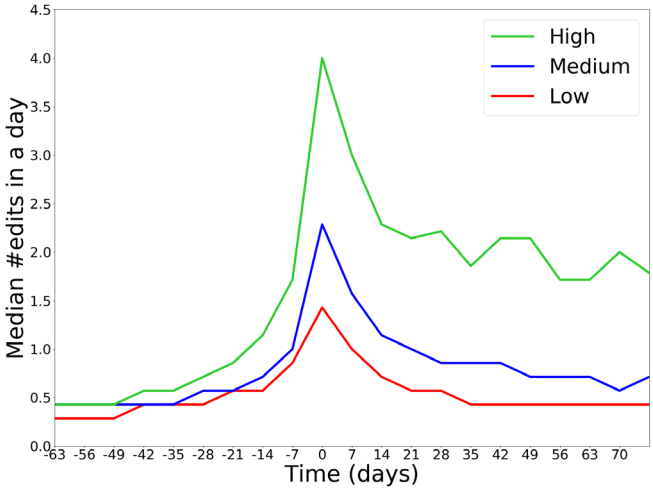

FIGURE 4 Median number of edits per day, centered around the day zero for achieving the silver badge [Color figure can be viewed at wileyonlinelibrary.com]

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.'s (2013) 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. Notice 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 analysis of variance (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).

5 | QUESTION 2: HOW DO LONG-TERM GROUP DYNAMICS CHANGE IN THE PRESENCE OF BADGES?

In this section, we explore 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, and so on).

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 5 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 Section 8, 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 6 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 from that of users who did not obtain the gold badge (Figure 5). 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

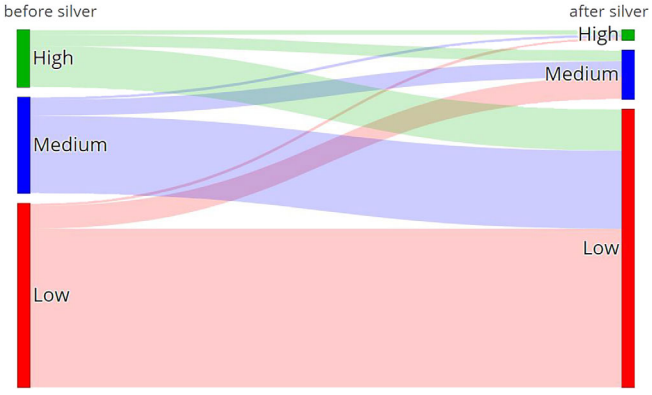

FIGURE 5 Flow between user groups before (left) and after (right) getting the silver badge [Color figure can be viewed at wileyonlinelibrary.com]

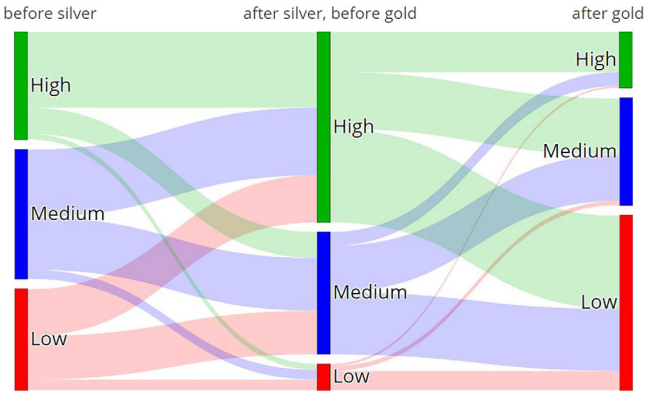

FIGURE 6 Flow between user groups before getting the silver badge (left), after getting the silver and before getting the gold badge (middle), and after getting the gold badge (right) [Color figure can be viewed at wileyonlinelibrary.com]

behavior identified by Anderson et al. (2013), and their activity levels decreased significantly.

5.1 | Are some badges gateways to other badge?

We now explore a different long-term dynamics on how two badges interact with one another. One badge is the edit badge, which we have explored so far, and the other, the voting badge, which we will delve into in Section 7.2. We use the available data on the period between September 2008 to May 2019 in SO.

Table 2 summarizes the amount of users in SO who obtained at least one of the following badges: edit-silver, edit-gold, vote-silver, or vote-gold. Not surprisingly, the number of users with vote badges is significantly larger than the amount of users who earned edit badges. This is, presumably, since voting badges are far easier to achieve: while a single edit requires entering a separate screen and making a change to the existing text, a vote requires only a single click. However, what is evident from the data is that achieving an edit badge is usually correlated with also achieving a vote badge. Indeed, 87% of the users who obtained the silver-edit badge also achieved the silver-vote badge. On the other hand, only 15% of the users who obtained the silver-vote badge also achieved the silver-edit badge.

When looking at the chronological order of obtaining these badges (Figure 7), some patterns emerge. The vast majority of the users begin by collecting the vote-silver badge. Out of the users who start from getting the silver-edit badge first, only a slight amount of users go on immediately for the gold-edit badge. Instead, most of them go toward achieving the silver-vote badge. Perhaps the gold-edit badge seems too hard to achieve at that point, so they choose a different badge as their next goal.

Focusing on the most productive and engaged population (in terms of editing and voting), this population includes 2,676 users who earned gold badges for both voting and editing action types. These users performed about 14.5 million votes out of all 138 million vote actions that were done on SO (about 10%). Amazingly, these users performed about 3,700,000 edits out of all 8,500,000 edit actions that were made on SO (about 43%!). As shown by Figure 7, the most popular path for achieving the four badges is in the following order: (a) silver-vote, (b) silver-edit, (c) gold-vote, and (d) gold-edit. These users first achieve all of the silver badges before moving on to achieving all of the gold badges. In addition, they almost always choose to achieve vote badges first.

Indeed, it seems these users are pursuing relatively short-term goals: choosing the action type that belongs to the most reachable badge in terms of its demands. According to this "algorithm," the users begin with the easiest action type (voting) and the closest badge (silver, by default). At this point, the users prefer to pursue the silver-edit badge rather than or the gold-vote badge, since the silver badge is more accessible for them. Moreover, it seems that the relatively easy to achieve silver-vote badge serves as a gateway to further badge achievements.

6 | QUESTION 2, CONTINUED: CAN WE MODEL USER BEHAVIOR?

Our machine learning model seeks to understand which users are susceptible to decreasing their work habits after obtaining the (silver) badge, so that we can target these users and tailor incentive solutions for them in order to increase their motivation (see Section 8). 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.

6.1 | 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.

TABLE 2 Amounts of users obtaining at least one of the four discussed badges in SO

		Edit badges		
		No badge	**Silver edit**	**Gold edit**
Vote badges	No badge		1992	96
	Silver vote	65,822	5,228	391
	Gold vote	11,091	5,934	2,676

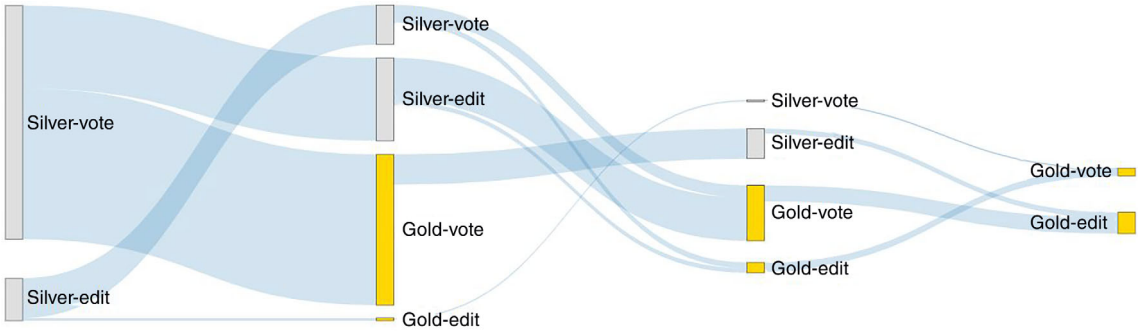

FIGURE 7 Chronological order of obtaining the edit and vote badges. People who received only a single badge (silver vote or silver edit) are not shown [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Prediction results for user decreasing contribution levels after day zero, using different combinations of features

Features	# features	Accuracy	F1 weighted	ROC AUC	Confusion matrix
User (U)	10	0.704	0.693	0.640	[7,987 1,549]
					[2,700 2,140]
Edits (E)	9	0.680	0.633	0.569	[8,680 856]
					[3,739 1,101]
Temporal (T)	40	0.835	0.839	0.875	[7,174 2,362]
					[124828]
U + E + T	59	0.837	0.842	0.875	[7,236 2,300]
					[424798]

Abbreviations: AUC, area under the curve; ROC, receiver operating characteristics.

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 $(c_1, ..., c_n)$ of n weeks of activity, these features average the consistency values $(c_1, ...c_3)$ for the first 3 weeks of activity and similarly for the first 5 and 10 weeks of activity prior to 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 $(c_{n-2}, ..., c_n)$ 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 relating to the intensity of users as well. 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.

6.2 | Prediction results

We used the XGBoost classifier algorithm (Chen & Guestrin, 2016) 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 10-fold cross-validation, with *SD* of results between runs smaller than 0.01

for all measures. As can be seen in Table 3, 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.

7 | QUESTION 3: DOES STEERING GENERALIZE?

In this section, we study whether steering generalizes to other SE projects and other badges.

7.1 | Generalizing to other SE projects

Figure 8 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: SO (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. (2013) 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.

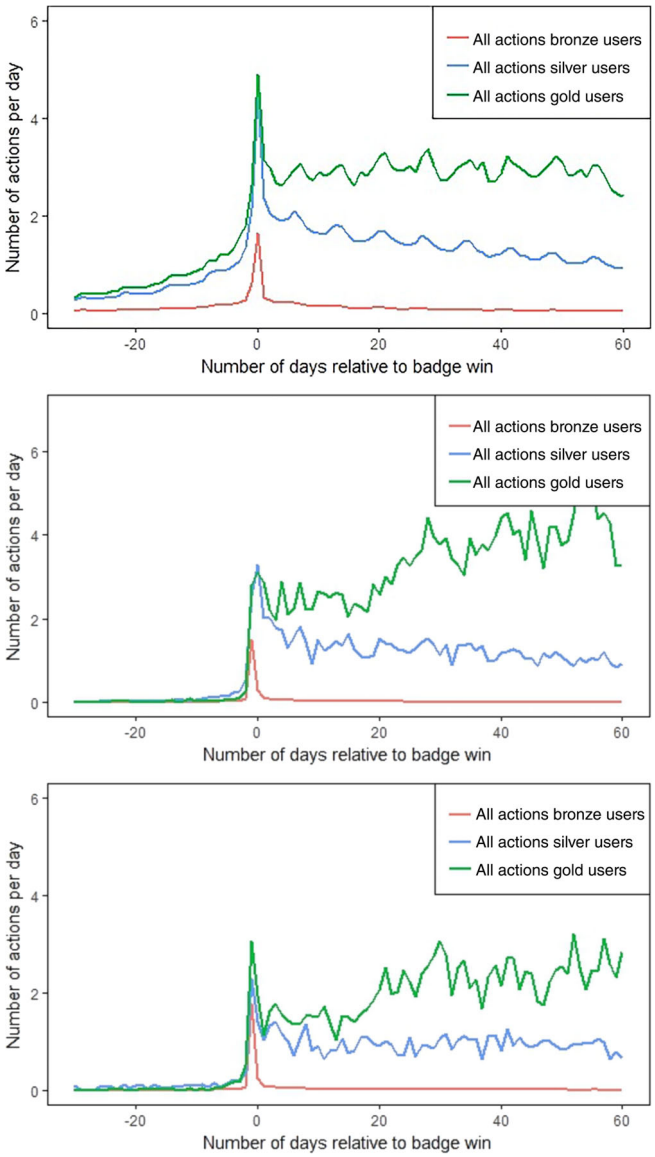

FIGURE 8 Average number of edit contributions as a function of distance from day zero for obtaining silver badge on the stack overflow (top), Ask Ubuntu (middle) and TeX-LaTeX SE sites (bottom) [Color figure can be viewed at wileyonlinelibrary.com]

Furthermore, as analyzed in Section 4, in these projects as well, the behavior of each badge group is different, and there is no "one-size-fits-all" phenomenon. The machine learning model that we have built for SO works in these projects as well and with similar performance.

7.2 | Generalizing to the vote action

As mentioned earlier, the two measures we defined in order to characterize user activity in the SE domain are work consistency and work intensity. We can generalize

these measures from dealing with edit actions to deal with other action types, including vote actions. This is the same activity that was analyzed by Anderson et al. (2014). We want to show how the analysis that dealt with the edit actions behavior can be generalized to vote actions, and point out the differences between behavior under both of these badge types. The data regarding the vote actions in each day was confidential and was achieved using a collaboration with the SO academic research department. The data are from the beginning of 2017, and we used all users who obtained the silver-vote badge after April 2017, so that we will have enough data for the time period before obtaining the badge.

We used the same consistency/intensity measures we used for the edit badges, and we use the k-means algorithm using the distance metric as described in Section 4.1 to cluster all users who got at least the silver-vote badge into three groups of activity: low, medium, and high. Figure 9 shows the clusters that were created, and Table 4 describes the amount of silver and gold users in each group of activity. The number of users in each cluster has changed dramatically from the edits analysis; here the largest group is of the medium activity and not the low as it was in the edit analysis. We can also see that the boundaries of the clusters "shifted," and now a user with a consistency median value of 0 and an intensity median value of 5 will be assigned to the low-activity group, rather to the medium-activity group in the edit analysis (same goes to the boundary between the medium- and high-activity groups). The reason for these changes is that voting is a much easier and more incidental action than editing.

As for the steering effect of the silver-vote badge, Figure 10 plots the contributions of the different engagement groups over time, relative to day zero, when the silver-vote badge was awarded. We observe the same phenomena as in Figure 4, regarding to how different user populations steer when receiving a badge. The high-activity users steer more than the other two groups, and they almost double the amount of votes in each day, comparing to their baseline activity prior to getting the badge. The other two groups of activity, medium and low, exhibit a far lower change in the activity on day zero and clearly return to their baseline activity. Results regarding the voting data are also significant.

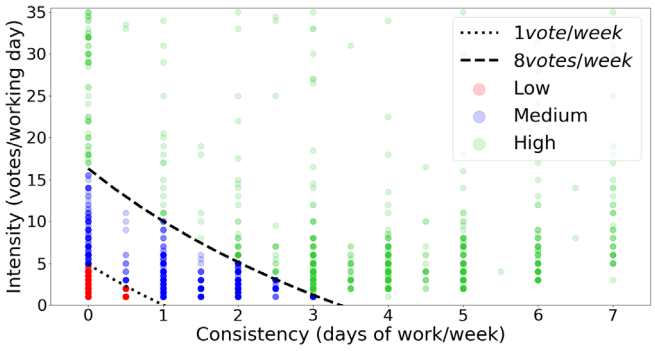

FIGURE 9 Scatter plot of user voting activity showing three user groups revealed by k-means ($K = 3$). Groups are distinguished using colors and boundary curves [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Number of gold-vote and silver-vote users in each activity group

	Low	Medium	High
Silver users	9,049	12,476	992
Gold users	368	993	960
Total	9,417	13,469	1,952

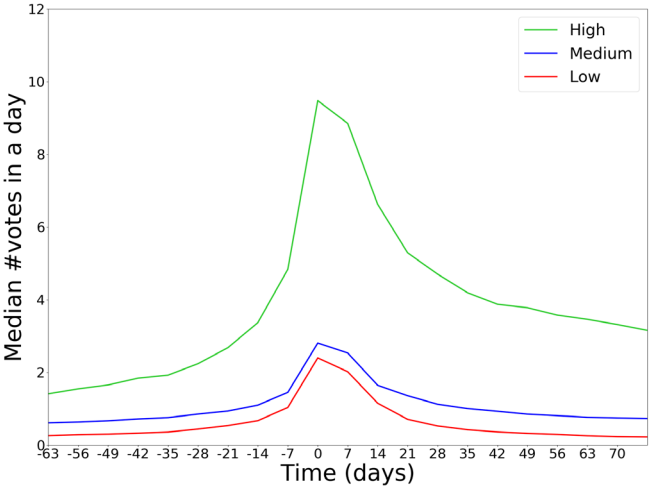

FIGURE 10 Median number of votes per day, centered around the day zero for achieving the silver-vote badge [Color figure can be viewed at wileyonlinelibrary.com]

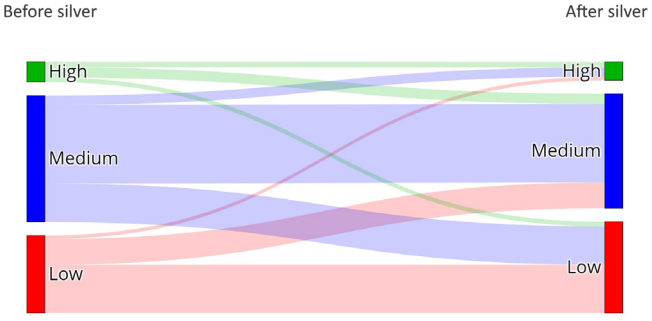

FIGURE 11 Flow between user groups before (left) and after (right) getting the silver-vote badge [Color figure can be viewed at wileyonlinelibrary.com]

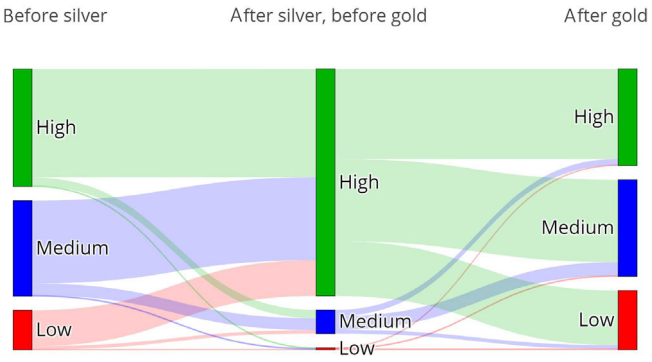

FIGURE 12 Flow between user groups before getting the silver-vote badge (left), after getting the silver-vote and before getting the gold-vote badge (middle), and after getting the gold-vote badge (right) [Color figure can be viewed at wileyonlinelibrary.com]

When looking at the users who received the silver-vote badge but not the gold-vote badge, we observe a different insight from the one we outlined for the same edit population. This group includes the users who made more than 300 votes in SO but less than 600 votes. In Figure 11, we track the flow of these users before and after obtaining the silver-vote badge. Here, unlike we have seen in Figure 5, the amounts of users in each group do not seem to change dramatically. Getting the silver-vote badge does not make users vote less than they did before getting it. On the other hand, getting the silver-vote badge also does not make them vote more than they did before. We can say that these users are indifferent to receiving the silver-vote badge.

When looking at the users who were able to get the gold-vote badge (Figure 12), we observe the insight we found regarding the gold-edit achievers (see Figure 6). Here the vast majority of the gold-vote users belongs to the high-activity group in the time period of between getting the silver-vote and gold-vote badges. Since the task of voting is easy for them, and they got some reward for doing it (the silver-vote badge), they highly increase their vote actions and become (or stay) high-activity users. For them, the silver-vote badge had a positive effect, it made them act more. Again, when the gold-vote badge is achieved, we observe a decrease in voting activity; the amount of users in each activity group is close to as it was before getting the silver-vote badge.

8 | DISCUSSION AND FUTURE WORK

Our main results in this article elaborate and expand previous research on badges (and in particular, Anderson et al. (2013)):

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 the 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.

One limitation of our approach is that the empirical analysis does not distinguish between steered and non-steered populations in SO. Indeed, motivations that underlie participants' activities in the site may be influenced by factors other than badge acquisition, such as motivation to contribute to the community. The "bump" in their activity can be explained by their natural activity patterns and would also occur in the absence of steering.

We answer this claim in two ways. First, we computed activity graphs for different acquisition thresholds for one of the badges in the study the voting ("Electorate") badge. The graph in Figure 13 shows the mean activity rate (count of vote actions per day) of participants as a function of time relative to some cumulative action threshold. The graph shows activity curves that are centered on the true Threshold (600 actions), Threshold − 10 (590 voting actions); Threshold + 10 (610 actions), and Threshold − 100 (500 actions).

The graph shows there can be significant differences in the contribution patterns depending on how the data are centered. The curve for Threshold + 10 is noticeably different from the other curves with the decline in contributions starting well before the Threshold cutoff point for this curve (the day in which 610 actions were completed). The spike of the Threshold + 10 curve is much lower than the other curves. We argue that this is due to participants decreasing their activity rates once the true threshold is crossed, which is in line with the steering phenomenon. The curves for Threshold − 10 and the true threshold are similar in that they show an increase of activity leading into the threshold cutoff followed by a sharp decline in activity. The spike of these curves is also similar, possibly because many people cross these two thresholds on the same day and get the badge. This trend is also in line with the steering phenomenon. The fact that the participation curves strongly depend on how the data are centered with respect to the badge threshold implies that at least part of their behavior can be attributed to the badge.

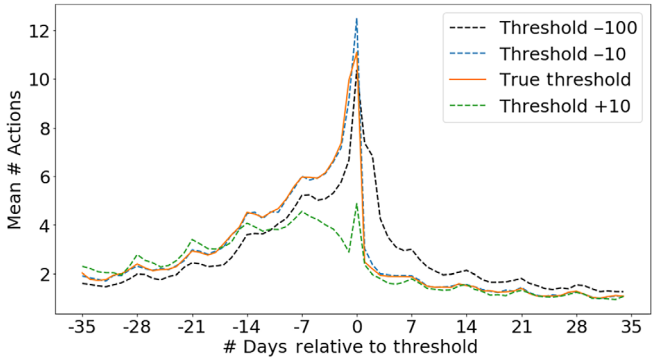

FIGURE 13 Mean activity as a function of time relative to threshold [Color figure can be viewed at wileyonlinelibrary.com]

Additional support for the effect of badges on behavior was provided by using the approach by Hoernle, Kehne, Procaccia, and Gal (2020) to separate those participants motivated by steering from those that do not deviate from their "default" activity distribution. Using this analysis, we discovered that about 38% of medium-activity users in SO (at least 8 edits per week, rarely work more than 4 days per week) are steered whereas only about 12% of high-activity (above 8 edits per week, work more than 3 days per week) users are steered. On the one hand, this shows that high level contributors will not be affected by better badge design, but these users are already strong contributors to SO. On the other hand, system designers can target medium-activity users, who are both regular contributors to SO and respond positively to badges.

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. [For qualitative badges, we may need to devise other measures for describing work progress towards a goal that go beyond consistency and intensity.] 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 Segal, Gal, Kamar, Horvitz, and Miller (2018).

ACKNOWLEDGMENTS

The authors would like to thank 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.

ENDNOTES

[1] For example, http://duolingo.wikia.com/wiki/Achievements

[2] For example, https://askubuntu.com/help/badges

[3] For example, https://blog.lyft.com/badge-glossary/

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

REFERENCES

Abramovich, S., Schunn, C., & 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., & Leskovec, J. (2013). Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 95–106). Rio de Janeiro, Brazil: ACM.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web* (pp. 687–698). Seoul, Korea: ACM.

Aoyagi, M. (2010). Information feedback in a dynamic tournament. *Games and Economic Behavior*, *70*(2), 242–260.

Bazelli, B., Hindle, A., & Stroulia, E. (2013). On the personality traits of stackoverflow users. In *Proceedings of the 29th IEEE international conference on software maintenance* (pp. 460–463). Eindhoven, The Netherlands: IEEE.

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

Bosu, A., Corley, C. S., Heaton, D., Chatterji, D., Carver, J. C., & Kraft, N. A. (2013). Building reputation in stackoverflow: An empirical investigation. In *Proceedings of the 10th Working Conference on Mining Software Repositories (MSR)* (pp. 89–92). San Francisco, CA: IEEE.

Cavusoglu, H., Li, Z., & Huang, K.-W. (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* (pp. 171–174). New York, NY: ACM.

Charleer, S., Klerkx, J., Odriozola, S., Luis, J., & Duval, E. (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, pp. 69–81). Paphos, Cyprus: CEUR-WS.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). San Francisco, CA: ACM.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, *71*(1), 1–27.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification: using game-design elements in non-gaming contexts. In *CHI '11: extended abstracts on human factors in computing systems* (pp. 2425–2428). Vancouver, Canada: ACM.

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

Grant, S., & Betts, B. (2013). Encouraging user behaviour with achievements: An empirical study. In *In Proceedings of the 10th Working Conference on Mining Software Repositories (MSR)*. San Francisco: California.

Hakulinen, L., Auvinen, T., & Korhonen, A. (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), 18–29.

Halavais, A., Kwon, K. H., Havener, S., & Striker, J. (2014). Badges of friendship: Social influence and badge acquisition on stack overflow. In *2014 47th Hawaii International Conference on System Sciences* (pp. 1607–1615). Waikoloa, Hawaii: IEEE.

Hamari, J., Huotari, K., and Tolvanen, J. (2015). Gamification and economics. In S. P. Walz & S. Deterding (Eds.), *The gameful world: Approaches, issues, applications*. Cambridge, Massachusetts: MIT Press.

Hickey, D. T., Willis, J., & Quick, J. (2015). Where badges work better. In *EDUCAUSE Learning Initiative* (Vol. 31, p. 2017). Washington, DC: EDUCAUSE.

Hoernle, N., Kehne, G., Procaccia, A. D., & Gal, K. (2020). The goal-gradient hypothesis in stack overflow. *arXiv Preprint* arXiv:2002.06160.

Immorlica, N., Stoddard, G., & Syrgkanis, V. (2015). Social status and badge design. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, May 18–22, 2015* (pp. 473–483). Florence, Italy: *ACM*.

Ipeirotis, P. G., & Gabrilovich, E. (2014). Quizz: targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on World wide web* (pp. 143–154). Seoul, Korea: ACM.

Jain, S., Chen, Y., & Parkes, D. C. (2009). Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM conference on electronic commerce (EC)* (pp. 129–138). California: Palo Alto.

Jia, Y., Xu, B., Karanam, Y., & Voida, S. (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* (pp. 2001–2013). San Jose, California: ACM.

Kobren, A., Tan, C. H., Ipeirotis, P., & Gabrilovich, E. (2015). Getting more for less: Optimized crowdsourcing with dynamic tasks and goals. In *Proceedings of the 24th international conference on world wide web* (pp. 592–602). Florence, Italy: ACM.

Li, Z., Huang, K.-W., & Cavusoglu, H. (2012). Quantifying the impact of badges on user engagement in online q&a communities. In *Proceedings of the International Conference on Information Systems (ICIS)* (pp. 3798–3807). Orlando, Florida: AIS.

Linehan, C., Kirman, B., & Roche, B. (2015). Gamification as behavioral psychology. In S. P. Walz & S. Deterding (Eds.), *The gameful world: Approaches, issues, applications*, Cambridge, Massachusetts: MIT Press.

Papoutoglou, M., Kapitsaki, G. M., & Mittas, N. (2018). Linking personality traits and interpersonal skills to gamification awards. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 214–221). Prague, Czech Republic: IEEE.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31.

Segal, A., Gal, K., Kamar, E., Horvitz, E., & Miller, G. (2018). Optimizing interventions via offline policy evaluation: Studies in citizen science. In *Proceedings of the 32nd AAAI conference on*

artificial intelligence (AAAI) (pp. 1536–1544). Louisiana: New Orleans.

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web* (pp. 1049–1054). Seoul, Korea: ACM.

Zhang, J., Kong, X., & Yu, P. S. (2016). Badge system analysis and design. *arXiv Preprint* arXiv:1607.00537.

How to cite this article: Yanovsky S, Hoernle N, Lev O, Gal K. One size does not fit all: A study of badge behavior in stack overflow. *J Assoc Inf Sci Technol*. 2021;72:331–345. https://doi.org/10.1002/asi.24409