Knowledge and Information Systems (2022) 64:475–506
https://doi.org/10.1007/s10115-021-01637-6

REGULAR PAPER

# The phantom steering effect in Q&A websites

Nicholas Hoernle[1] · Gregory Kehne[2] · Ariel D. Procaccia[2] · Kobi Gal[1,3]

Received: 12 March 2021 / Revised: 3 December 2021 / Accepted: 3 December 2021 /
Published online: 8 January 2022

**Abstract**
Virtual rewards, such as badges, are commonly used in online platforms as incentives for promoting contributions from a userbase. It is widely accepted that such rewards "steer" people's behaviour towards increasing their rate of contributions before obtaining the reward. This paper provides a new probabilistic model of user behaviour in the presence of threshold rewards, such a badges. We find, surprisingly, that while steering does affect a minority of the population, the majority of users do not change their behaviour around the achievement of these virtual rewards. In particular, we find that only approximately 5–30% of Stack Overflow users who achieve the rewards appear to respond to the incentives. This result is based on the analysis of thousands of users' activity patterns before and after they achieve the reward. Our conclusion is that the phenomenon of steering is less common than has previously been claimed. We identify a statistical phenomenon, termed "Phantom Steering", that can account for the interaction data of the users who do not respond to the reward. The presence of phantom steering may have contributed to some previous conclusions about the ubiquity of steering. We conduct a qualitative survey of the users on Stack Overflow which supports our results, suggesting that the motivating factors behind user behaviour are complex, and that some of the online incentives used in Stack Overflow may not be solely responsible for changes in users' contribution rates.

**Keywords** Virtual badges · Privileges · Steering · Goal-gradient hypothesis · Amortised inference

## 1 Introduction

A well-known finding from behavioural science research is that efforts towards a goal increase with proximity to that goal. This phenomenon, termed the goal-gradient hypothesis, has been demonstrated in a variety of settings, from animal studies in the laboratory to consumer purchasing behaviour [12,20]. More recently, the goal-gradient effect was observed in people's

✉ Nicholas Hoernle
s1841130@ed.ac.uk

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

[2] Harvard University, Cambridge, USA

[3] Ben-Gurion University, Beersheba, Israel

behaviour in online communities that use virtual rewards, such as badges and reputation points, to increase users' contributions to the site [1,24]. In these contexts, the goal-gradient hypothesis has been referred to as "steering" [1,2]. Recent examples of online settings that use badges include communication platforms such as MS teams, ride-sharing platforms such as Lyft and online learning platforms such as Duolingo.

In this paper, we study the steering phenomenon, in one such community, that of Stack Overflow (SO), where users can acquire badges and obtain reputation points for making different contributions to the platform, such as editing or voting on posts. We identify *who* exhibits steering, who does not, and *how* this steering behaviour can be characterised from observational data. Our surprising result is that a large population (at least 60%) of badge achievers, who are highly active, do not appear to exhibit steering towards those badges.

We present a generative model of steering which models users as having default activity rates which they can deviate from when approaching the requirements for achieving a badge. The model is able to fit a complex multimodal distribution over the parameters that govern users' activities. This allows it to capture different levels of steering in the population. We applied the model to data collected from thousands of SO users, and investigate the following research questions:

1. Are all badge achievers affected by the steering (or goal-gradient) hypothesis in the same way?
2. If some users do not steer, what portion of the population falls under this category?
3. Does the presence of these users in the dataset change any conclusions that were previously drawn about the phenomenon of steering?

Our results revealed the following insights: First, more than 60% of the users are not steered, in that they exhibit a consistent activity rate in SO that is not effected by the badge. We prove that a "bump" in activity that is conveyed by prior work arises as an artifact of centering the data on the day of badge achievement [1,28]. We call this phenomenon *the Phantom Steering Effect*. Second, about 5–30% of users are steered, in that they dramatically increase their rate of activity prior to achieving the badge. It is the effect that this small population of steered users has on aggregate measures that has led to the previous and broader claims of steering [1,24,28]. Third, a large portion of these steered users decrease their activity rate beyond what is claimed in prior work [1], reaching close to 0 after the badge has been achieved. Our conclusions are supported by responses to a user survey that included 70 active SO participants, in which only 24% of participants selected badges as a motivating factor for their contribution.

We extend our approach to modelling people's behaviour under another popular incentive mechanism in SO, that of reputation points thresholds. When users cross predefined thresholds, they earn privileges on the site. For example, crossing 200 points results in a reduction of advertisements; $1K$ points denotes users as "established" and gives them the option to see the total count of both up and down votes on a post; and $20K$ points unlocks further editing, deletion and un-deletion privileges. There are other thresholds, all associated with privileges on SO that can be found on the SO webpage.[1] We argue that crossing a threshold, and earning the associated privileges, can be viewed in the same light as earning a badge [13]. Thus, in this work, we occasionally refer to the achievement of a reputation threshold and the achievement of a badge synonymously. This investigation applies the same model used for badges to the reputation point threshold and investigates whether the above hypotheses hold in this new setting.

[1] https://stackoverflow.com/help/privileges/.

Our results revealed that more than 90% of the threshold achievers were not steered by the threshold. For the small minority of users that did change their behaviour, this change mirrored that from the badge study. Moreover, we find an inconsistency between the qualitative, self-reported results from the user-survey and the computational results that are presented. Users claimed that the privileges were a motivating factor towards further contribution to SO but our computational results suggest that the effect is limited. As such, we posit that such rewards may still contribute towards an ecosystem that can keep users engaged even if the goal gradient effect is not directly displayed.

Our study has important ramifications for system designers who invest resources into the implementation of badge rewards systems and for researchers who wish to understand the factors that contribute towards users' continued participation in online communities. It provides a sobering perspective on the efficacy of badges and reputation point thresholds as effective incentives, in that for much of the population, the steering effect does not appear to hold. This does not mean that the ecosystem fails to incentivise users. It is possible that rewards that foster a "sense of community" [13] that engages users toward continued contribution. However, our results do suggest that the steering effect (goal gradient hypothesis) holds only in a limited capacity [1,24].

## 2 Related work

We begin by relating to the general literature on the effect of badges in online communities. We then present in detail the specific work of Anderson et al. [1] which helps to motivate the generative models that we develop in Sect. 3.

### 2.1 The study of online badges

The goal-gradient hypothesis stems from behavioural research where animals were observed to increase their effort as they approach a reward [12,20]. Kivetz et al. [20] studied the behaviour of different populations of people who were working towards various rewards. They concluded that the goal-gradient hypothesis also holds true for people. Subjects who received a loyalty card, which tracked the number of coffees purchased from a local coffee chain, purchased coffee significantly more frequently the closer they were to earning a free cup of coffee. The authors recognised the existence of a group of participants who did not complete their coffee cards for the duration of the study and did not exhibit a noticeable change in their coffee purchasing habits. They concluded that the loyalty card effect was constrained to the population of participants who handed in their completed loyalty cards in exchange for the free-coffee reward. However, the authors had no means for estimating what fraction of users did not submit their cards and therefore they could not estimate how pervasive this effect might be when evaluated on the population at large.

Anderson et al. [1] and Mutter and Kundisch [24] were the first to study the goal-gradient hypothesis in online settings. They studied the *observed* effect of badges on the behaviour of participants in large Q&A sites. Both studies found evidence that users increase their rate of work as they approach the badge threshold. However, they did not address the possibility that some users might achieve the badge as a consequence of their routine interactions on the website rather than being steered by the badge. There is a possibility that people's actions are governed by motivations other than badges. We extend these works by allowing for this possibility, such that we can characterise the true changes to users' behaviour under the influence of a badge and distinguish this from the case where users do not noticeably change their interaction behaviour.

Other studies have independently confirmed that the presence of online badges increases the probability that a user will act in a manner to achieve the badge, as well as the rate at which the user will perform those actions [6,14,21,28]. Kusmierczyk and Gomez-Rodriguez [21] highlight the importance of modelling the "utility heterogeneity" among the users but they study badges which have a threshold of 1 action and do not characterise *how* one might change one's behaviour in the presence of the badge incentive. Yanovsky et al. [28] study the presence of different populations within the SO database by employing a clustering routine. They discovered notably different responses to the badge based on the cluster that a user belongs to. Their study did not acknowledge the possibility that the observed data might be consistent with a hypothesis that some users do not exhibit steering. Anderson et al. [2] studied the implementation of a badge system in a massive open online course and they provide a prescriptive system for the design of badges such that there is a maximum effect on the population. Zhang et al. [29] suggest that SO create new badges to encourage users to integrate helpful comments into the accepted answers. They thereby present an example of how system designers might use a badge to encourage a desired behaviour from their user base. In contrast to this, we suggest that badges have a limited scope and work should be completed to understand other motivations that the users' have such that better and more effective rewards can be designed to motivate online communities.

### 2.2 A utility model for steering

Most relevant to our work is the paper from Anderson et al. [1] who present a parametric description of a user's utility when the user is steered by badges. The model describes users as having their own preferred distribution from which actions are sampled. As users approach the required threshold for achieving a badge, they *deviate* from their preferred distributions. The deviation from the preferred distribution is controlled by the utility gained by achieving the badge and the cost for deviating from the preferred distribution.

We let $A_u^d$ refer to the distribution over the count of actions that a user $u$ takes on day $d$. The user's utility is a function of $A_u^d$ and it is the sum of three terms.[2] The first term, $\sum_{b \in B} I_b V_b$, is the non-negative value that a user derives from already-attained badge rewards (where $V_b$ is the assumed value of a badge and $I_b$ is the indicator that the user has attained badge $b$). The second term, $\theta \mathbb{E}_{A_u^d}[U_{u,d+1}(A_u^{d+1})]$, describes the user's expected future utility, discounted by $\theta$, when acting under the distribution $A_u^d$. The final term, $g(A_u^d, P_u^d)$, is a cost function that penalises the user for deviating from the preferred distribution $P_u^d$ on that day. The cost $g$ represents the unwillingness of the users to change their behaviour, and it is in tension with the users' desire to achieve future badges.

The utility on day $d$ for user $u$ is then [1]:

$$U_{u,d}(A_u^d) = \sum_{b \in B} I_b V_u^b + \theta \mathbb{E}_{A_u^d} \left[ U_{u,d+1} \left( A_u^{d+1} \right) \right] - g \left( A_u^d, P_u^d \right)$$

[2] Our notation differs slightly from that of Anderson et al. [1]. Anderson et al. [1] uses a parameter $\mathbf{x_a}$ to refer to a user's distribution over the next action. We rather use $A_u^d$ to denote the distribution over the count of actions on a particular day. The two are linked (the distribution over the next action influences the count of actions on a specific day); however, we choose to model directly the data that is available from SO rather than a quantity that we do not observe.

It is important to note that the cost term $g$ is non-zero only when users deviate from their preferred distribution $P_u^d$. As such, this model assumes users deviate only to attain the value from the badge and only if that value outweighs the cost that is paid for deviating. This means that a deviation on the rate of actions which are incentivised by the badge must be an increase before the badge is achieved and cannot be an increase after the badge is achieved (under a standard utility-theoretic assumption that all the utility of the badge is conveyed to the user upon receipt of the badge). We will make these same assumptions in the models presented in Sect. 3.1.

This utility-based model presents a compelling description of how people respond to badges; however, it was not evaluated or tested by fitting it to specific data from SO. Rather, predictions of the model were compared to aggregated data from SO and we show in Sect. 7 that the aggregated analysis from these count data can lead to incorrect conclusions. The lack of analysis on individual-level predictions limits the credibility of the study as well as its practical value—it is difficult to apply the utility-based model to the mechanism design problem of badge placement without a means for determining the appropriate model parameters for a given community of contributors.

In this work, we address the shortcomings of the utility-based approach by introducing a probabilistic model which allows us to use the vast literature on posterior inference in such models to assist with parameter estimation [4,17,19,26,27]. The probabilistic model has two advantages over this prior work: (1) posterior distributions for latent parameters in the model can be learnt from real-world interaction data and (2) the model's fit to data can be used to test and update scientific hypotheses (for example, in this paper we propose and validate that while some users may steer in a similar way, there exist users who do not experience steering).

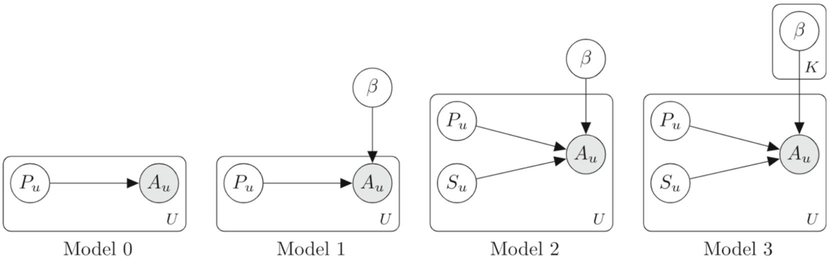
## 3 Modelling user activities

We model users' activities in SO as a distribution over their action counts. The model aims to incorporate the major aspects of the utility model from Anderson et al. [1] but it frames the problem such that parameters can be estimated from data and the models can be tested on their fit to unseen user action data to allow for model comparison [4,7]. Moreover, the model allows for different users to experience different levels of steering allowing for a more detailed investigation into the steering phenomenon.

### 3.1 A generative model of steering

Let $P_u$ be a latent parameter that controls the rate of activity for user $u$; this is the *preferred distribution* of user $u$. $P_u$ induces a probability distribution over the action counts $A_u$ of user $u$. Let $\beta$ denote the deviation of the user's activity from $P_u$ as a result of steering. The observed data for each user, $A_u$, consists of daily action counts for a predetermined number of weeks before and after achieving the badge. Thus, for $D$ days of interaction, $P_u$, $A_u$ and $\beta$ are all vectors of length $D$.

Figure 1 presents four plausible generative models of user behaviour in SO where each model presents an increasingly complex description of how people might respond to badge incentives. White circles denote latent random variables and colored circles denote observed random variables; solid lines represent conditional dependence between the random variables. Model 0 (Fig. 1, left) describes a non-steering scenario, in which the observed action counts,



[htb]

**Fig. 1** Model 0 (baseline model) has no notion of a badge—only a user's preferred distribution induces the distribution over the observed actions. Model 1 allows for a global badge deviation ($\beta$) from a user's preferred distribution and it is experienced by all users. Model 2 has a user-specific strength parameter ($S_u$) that selects whether or not user $u$ adheres to $\beta$. Model 3 allows for multiple parameters ($\beta^k$, $k \in 1, \ldots, K$) that the users might adhere to, in this model ($S_u$) becomes a switching variable that chooses between the $\beta^k$ parameters

$A_u$, depend only on the user's preferred distribution, $P_u$. Model 1 (Fig. 1, center-left) is a steering model in which all users deviate systematically from $P_u$ in a manner that is controlled by $\beta$. As the values for $\beta$ increase (above 0), the users experience an increased activity rate (above their preferred distribution). Similarly, as $\beta$ decreases (below 0), the users experience a decreased activity rate. Model 1 assumes that all users are steered in the same way. Model 2 (Fig. 1, center-right) relaxes this assumption by introducing a user-specific Bernoulli parameter $S_u \in \{0, 1\}$ dictating whether or not user $u$ adheres to the effect of $\beta$. Finally, we introduce Model 3 (Fig. 1, right) which allows for $K$ different deviations where each deviation, $\beta^k$, describes a different response to the badge incentive. For this model, $S_u \in \{0, \ldots, K\}$ now represents a Categorical random variable that selects which deviation, $\beta^k$, that user $u$ adheres to.

The parameter $\beta$, that controls how a user responds to a badge, is a vector of length $D$ (each day relative to the date of badge achievement). Reflecting the intuition that steering positively influences a user before the badge achievement, we constrain $\beta$ to be strictly positive before *day* 0—the day when the user achieves the badge. Moreover, for the Models 1 and 2, we constrain $\beta$ to be strictly negative after this day to reflect the intuition that a user gains no further utility from the badge once it has been achieved (and thus does not work harder than his preferred distribution $P_u$). $\beta$ therefore implicitly includes the trade-off between the cost function $g$ and the badge utility $V$ that is discussed in Sect. 2.2. We relax this second constraint for Model 3 to test the hypothesis that users maintain their base rate of activity well after the achievement of the badge, as is described by Anderson et al. [1], Yanovsky et al. [28].

Model 3 includes three possibilities for $\beta^k; k \in \{1, 2, 3\}$. $\beta^1$ sets the deviation to $\mathbf{0}$ implying no deviation, and capturing the assumptions of Model 0. $\beta^2$ uses the same assumptions from the Models 1 and 2 above in that $\beta^2$ is strictly positive before the badge is achieved and strictly negative after this day. Finally, $\beta^3$ is strictly positive before the badge is achieved but it is set to $\mathbf{0}$ after this day. These details are summarised in Table 1.

### 3.2 Likelihood of action counts

In this section, we define the parameters that govern the distribution over users' action counts in SO. We wish to describe a variety of behaviours, including users who contribute

**Table 1** Table detailing the constraints on the $\beta^k$ parameters and which models these parameters apply to

| | Deviation from $P_u$ | Model 0 | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|---|
| $\beta^1$ | Set to **0**; No deviation | ✓ | ✓ | ✓ | ✓ |
| $\beta^2$ | Non-negative before badge; Non-positive after badge | | ✓ | ✓ | ✓ |
| $\beta^3$ | Non-negative before badge; **0** after badge | | | | ✓ |

sporadically and those who are more consistent. We therefore model action counts using a zero-inflated Poisson distribution. The zero-inflated Poisson distribution has a rate parameter, $\lambda_u^d$, and a Bernoulli probability, $\alpha_u^d$, associated with each user $u$ and each day $d$ of interaction. The Bernoulli parameter, $\alpha_u^d$, describes the probability that user $u$ is active or not on a given day $d$. The rate parameter, $\lambda_u^d$, describes the expected count of actions that the user will perform under a Poisson distribution, conditioned on the user being active on the platform. Note that a user can be active on the platform without performing an action (e.g. logs on to the SO website but does not contribute). Conceptually, this would correspond to drawing a 1 from the Bernoulli distribution but a count of 0 actions from the Poisson distribution.

The probability that user $u$ performs $m$ actions on day $d$ is presented in (1). We refer to the parameters $\alpha_u^d$ and $\lambda_u^d$ as a user's *rate parameters* for day $d$.

$$Pr[A_u^d = m] = \begin{cases} (1 - \alpha_u^d) + \alpha_u^d Pois(0 \mid \lambda_u^d), & \text{if } m = 0 \\ \alpha_u^d Pois(m \mid \lambda_u^d), & \text{otherwise} \end{cases} \tag{1}$$

### 3.3 Deriving the rate parameters $\alpha_u$ and $\lambda_u$

This section connects the rate parameters, $\alpha_u$ and $\lambda_u$, to the generative models of Sect. 3.1. Each of $P_u$, $\beta^k$ and $S_u$ includes one component for $\alpha_u$ and one component for $\lambda_u$. As such, for $D$ days of interaction, $P_u = (P_{u,\alpha}, P_{u,\lambda})$ comprises two real-valued vectors, each of length $D$. $P_{u,\alpha}$ is the user's preferred distribution that is associated with $\alpha_u$ and $P_{u,\lambda}$ is the user's preferred distribution associated with the parameter $\lambda_u$. Similarly, $\beta^k = (\beta^k_\alpha, \beta^k_\lambda)$ comprises two real-valued vectors of length $D$ that are associated with $\alpha_u$ and $\lambda_u$, respectively. Finally, $S_u$ is a tuple of two Categorical variables (of order $K$) that selects among the steering parameters $\beta^k$. When there is only one steering parameter, Model 2 is accurately described by Model 3 by setting $K = 2$ and $\beta^1 := \mathbf{0}$. In this special case, the variable $S_u$ becomes a Bernoulli random variable that indicates the presence (or lack thereof) of the steering parameter $\beta^2$. As such Model 2 is a simplification of Model 3; similarly, Models 0 and 1 can be seen to simplify Model 2.

Equation (2) derives a vector of probability values $\alpha_u$ (one for each day of interaction) as the element-wise sigmoid transformation of a vector that is the addition of the user's preferred distribution $P_{u,\alpha}$ with $\beta^j_\alpha$ where $\beta^j_\alpha$ is the steering parameter that is selected by

$S_{u,\alpha} = j$. Equation (2) also derives a vector of strictly positive rate values $\lambda_u$ (one for each day of interaction) as the element-wise softplus transformation of the vector $P_{u,\lambda} + S^j_{u,\lambda} \times \beta^j_\lambda$. Below, $\mathbb{1}^j$ refers to the indicator variable that is 1 if $S_u, . = j$ and 0 otherwise.

$$\begin{aligned} \alpha_u &= \sigma\left(P_{u,\alpha} + \sum_{j=1}^{K} \mathbb{1}^j_{S_{u,\alpha}} \times \beta^k_\alpha\right) \\ \lambda_u &= softplus\left(P_{u,\lambda} + \sum_{j=1}^{K} \mathbb{1}^j_{S_{u,\lambda}} \times \beta^k_\lambda\right) \end{aligned} \tag{2}$$

The complete generative description for Model 3 is as follows (Models 2, 1 and 0 are generated in the same way with the corresponding restrictions on $\beta$ and $S_u$):

1. Sample $P_u$ and $S_u$ from their prior distributions (see Sect. 4).
2. Compute $\alpha_u$ and $\lambda_u$ using (2).
3. Sample the vector of the counts of actions for user $u$ from the zero-inflated Poisson likelihood as in (1).

In practice, we wish to model the activity of users on SO as they progress through time as accurately as possible. We therefore employ a recurrent model, and in our experiments we used a GRU with a single hidden layer [10]. This approach uses the product rule of probability to factor the joint distribution over actions through time (recalling that the actions also depend on the users' steering parameters, $S_u$, and their preferred distributions, $P_u$). Below the notation $A_{u,<T}$ is used to refer to all actions users $u$ performs before time step $T$:

$$\begin{aligned} p(A_{u,\leq T} \mid P_u, S_u) = p(A_{u,T} \mid A_{u,<T}, P_u, S_u) \dots \\ p(A_{u,T-n} \mid A_{u,<T-n}, P_u, S_u) \dots p(A_{u,1}, P_u, S_u) \end{aligned} \tag{3}$$

## 4 Amortised variational inference for steering

A fully-specified generative model defines a joint distribution over some latent random variables, $P_u$ and $S_u$, and the observed random variables, $A_u$. The challenge is to infer the posterior of the latent parameters given the data that were actually observed: $p(P_u, S_u \mid A_u)$. For all but a handful of conjugate models, the posterior is intractable to derive analytically [5,11,25].

Rather, to infer the underlying parameters in the latent space, we use amortised variational inference [17,18,26]. Amortised inference uses a neural network to encode a data point into the latent parameters that are associated with its approximate posterior distribution. Moreover, the inference objective allows model comparison such that hypotheses about the data can be tested (e.g. allowing us to validate the inclusion of the steering parameter, $S_u$).

Variational inference is a popular method for approximating the intractable posterior distribution by introducing a different (and more easily sampled from and evaluated) distribution over the same latent variables: $q(A_u, S_u) = q(A_u)q(S_u)$. By minimising the KL-divergence between $q(A_u, S_u)$ and the true posterior $p(A_u, S_u \mid A_u)$, one obtains an approximation to the true posterior [11].

It is important to note that minimising the KL-divergence between $q(A_u, S_u)$ and $p(A_u, S_u \mid A_u)$ is equivalent to maximising the variational objective, called the Evidence Lower BOund (see Hoffman et al. [11] for a derivation and discussion of the ELBO). This *ELBO* derives its namesake from the fact that it lower-bounds the marginal log-likelihood of the data under the assumptions of the model, a fact easily derived in Equations 4 and 5, where Jensen's inequality is applied in the second line of Equation 5 [3]. It is due to this lower bound on the marginal log-likelihood, that it is also common to use the ELBO for model comparison, as is done in Sect. 6.1 [8].

$$\begin{aligned}\log p(A_u) &= \log \int \sum_{S_u} p(A_u, P_u, S_u) \partial P_u \\ &= \log \int \sum_{S_u} q(P_u, S_u) \frac{p(A_u, P_u, S_u)}{q(P_u, S_u)} \partial P_u\end{aligned} \tag{4}$$

The second line in Eq. 4 can be recognised as computing the expectation of $\frac{p(A_u,P_u,S_u)}{q(P_u,S_u)}$ with respect to the approximating distributions $q(P_u, S_u) = q(P_u)q(S_u)$. Moreover, we assume $q(P_u)$ exists in a distributional family where it is possible to compute the pathwise derivative via the reparameterisation trick [17]. As the steering parameter, $S_u$, is not continuous, this same reparameterisation cannot be done. It is possible to replace the Categorical variable with a continuous approximation as is done by Maddison et al. [23] and Jang et al. [15]; or, if the dimensionality of the Categorical variable is small, it can be marginalised out [18]. We choose this latter approach leading to the ELBO as defined in Eq. 5.

$$\begin{aligned}\log p(A_u) \;&=\; \log \mathbb{E}_{q(P_u,S_u)} \left[ \frac{p(A_u, P_u, S_u)}{q(P_u, S_u)} \right] \\ &\geq\; \mathbb{E}_{q(P_u,S_u)} \left[ \log \frac{p(A_u, P_u, S_u)}{q(P_u, S_u)} \right] \\ &=\; \sum_{S_u} \mathbb{E}_{q(P_u)} \left[ q(S_u)(\log p(A_u, P_u, S_u) - \log q(P_u) - \log q(S_u)) \right] \\ &:=\; ELBO(A_u)\end{aligned} \tag{5}$$

Following standard practice $q(P_u)$ is assumed to be an isotropic Gaussian with $\mu_\Phi(A_u)$ and $\sigma^2_\Phi(A_u)$ computed by an inference (encoding) network with parameters $\Phi$. The prior $p(P_u)$ is a standard normal Gaussian distribution. Similarly, the categorical encoding distribution $q(S_u)$ simply computes the probability that user $u$ belongs to class $j$, $j \in \{1, \ldots, K\}$.[3]

## 5 Data domains for empirical study

We consider two types of threshold rewards that are present on SO. The first is the threshold badge rewards that are awarded for completing common actions on the website. Completing the required action directly progresses a user towards the threshold for achieving the badge. The second type of threshold reward are the privileges that are awarded for reaching a predefined number of reputation points. These privileges "*control what [users] can do on Stack Overflow [and users] gain more privileges by increasing their reputation*."[4] The privilege rewards are in contrast to the badge rewards that we study in that the reputation point system requires feedback from other users, in the form of accepts and upvotes, whereas a user can progress towards a threshold badge directly by completing the requisite action [1]. We aim

[3] All modelling and inference code can be found at the repository: https://github.com/NickHoernle/icdm2020.

[4] https://stackoverflow.com/help/privileges.

**Table 2** Table detailing the badge rewards under study

| Badge | Incentivised action | Threshold | # Users |
|---|---|---|---|
| Electorate | Votes on questions | 600 | 5701 |
| Civic duty | Votes on questions and answers | 300 | 20,880 |
| Copy editor | Edits | 500 | 750 |
| Strunk and White | Edits | 80 | 3101 |

to investigate the prevalence of steering in these two settings and to document any structural differences in how people respond to these different reward types.

We consider four common badge types on SO. Table 2 details the different badges that we study. We present: **Incentivised Action**—the specific action(s) that the badge is designed to incentivise; **Threshold**—the required number of that actions that should be completed to achieve the badge; and, **# Users**—the number of users in the sample that have achieved the badge. Note that the Electorate badge incentivises one of the same actions (question-votes) as the Civic Duty badge but it has a higher requirement for achievement. We have removed all the users who achieved the Electorate badge from our study of the Civic Duty badge, to remove the confounding effect of the Electorate badge on the users who achieved Civic Duty. The same holds for the Copy Editor badge which incentivises the same action (edits) as the Civic Duty badge. Additional details can be found about these badges, and others, on the SO website.[5] In the event that more than one action is directly incentivised by the badge (e.g. for the Electorate badge), we model the combined activity by summing over the different action types. The interaction data were kindly supplied by SO in an anonymised form and it consists of the action counts per day of users on the website from January 2017 to April 2019.

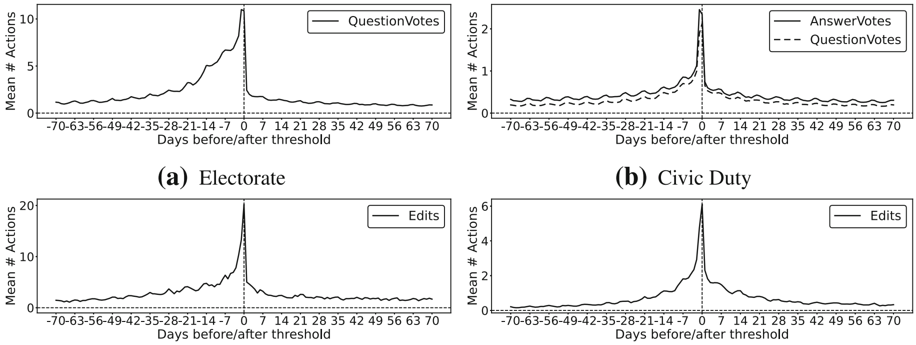
Figure 2 presents the mean number of actions per day averaged across the entire user base for 70 days before and 70 days after the users achieved the badge. We plot only the actions that are directly incentivised by the badge. The steering effect, as described by Anderson et al. [1] and Mutter and Kundisch [24], can clearly be seen by the increase in the rate of actions leading into the badge achievement date. After the badge has been achieved, the rate of activity rapidly drops and returns to a more constant rate of interaction [1]. The steering effect is most evident on the interaction data from the Electorate and Copy Editor users (Fig. 2a) but the same general increase and then decrease can be seen in the trends from the other badges.

Next we consider four different reputation point thresholds that unlock different privileges on SO. Users achieve reputation points on SO by completing a number of different actions and critically by having other users validate their contributions. For example, users achieve reputation points by having their questions and answers upvoted, by having their answers accepted or by having their edits accepted. Table 3 details the different thresholds for gaining privileges that we study. We present: **Threshold**—the required number of reputation points that should be achieved to unlock the privilege; **# Users**—the number of users in our dataset that have achieved the privilege; and, **Unlocked Privileges**—a brief description of the privileges that are unlocked on the website. Other reputation thresholds and their associated privileges can be found on the SO website.[6] The reputation data were obtained from the

[5] https://stackoverflow.com/help/badges.

[6] https://stackoverflow.com/help/privileges/.

[htb]



**(a)** Electorate

**(b)** Civic Duty

**(c)** Copy Editor

**(d)** Strunk & White

**Fig. 2** Plot of the mean count of actions per user per day 10 weeks before and 10 weeks after the users achieved the corresponding badges. Notice the different limits on the *y*-axis for the average number of actions that are performed

**Table 3** Table detailing the reputation privileges under study

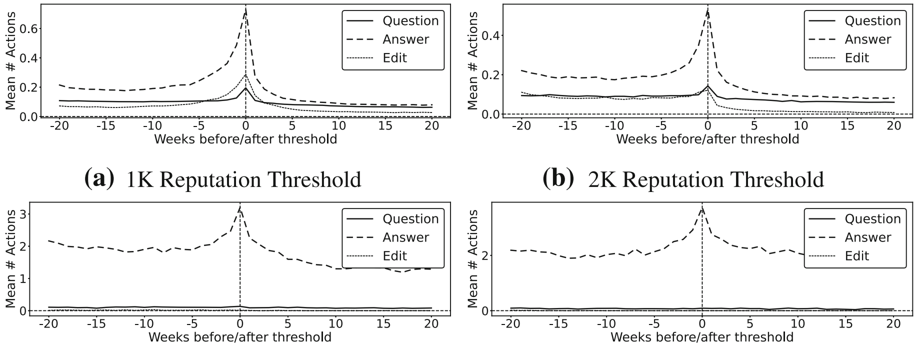
| Threshold | # Users | Unlocked privileges |
|---|---|---|
| 1K | 71,795 | Established user: view the vote counts on posts |
| 2K | 29,161 | Edit questions and answers: edits to posts are applied immediately without being reviewed |
| 20K | 1316 | Trusted user: expanding editing, deletion and un-deletion privileges |
| 25K | 966 | Access to site analytics: access to internal and Google site analytics |

publicly available SO data dump[7] and it was filtered to users who joined SO after 2012/01/01. We study the interaction data aggregated by week due to the sparsity of the actions through time.

Figure 3 shows the mean number of actions per week, averaged across users, for 20 weeks before and 20 weeks after crossing the defined reputation threshold. Differences in the rates of activity can be seen before and after the threshold was achieved; with a higher rate before the threshold and a lower rate after the threshold. Again, this appears to reflect the steering hypothesis—especially for the lower thresholds. Moreover, different behaviours around the different reputation thresholds are evident. The rates of answering are much lower for the lower thresholds than for the higher thresholds.

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

[htb]



**(a)** 1K Reputation Threshold **(b)** 2K Reputation Threshold

**(c)** 20K Reputation Threshold **(d)** 25K Reputation Threshold

**Fig. 3** Plot of the mean count of actions per user per day 10 weeks before and 10 weeks after the users achieved the corresponding reputation threshold

A further point of interest is evident in Fig. 3b: The rate of editing from these users decreases to near 0 after the badge has been achieved. This is in comparison with the $1K$ threshold where the change in editing behaviour appears symmetric around the origin and the $20K$ and $25K$ thresholds where this rate is consistently low. A plausible reason for this is that once a user crosses the $2K$ reputation threshold, they no longer receive reputation points for editing posts.[8] This provides evidence that, for some users, the points that they receive for editing do serve to motivate their contributions.

## 6 Empirical study

We begin by detailing the evaluation criteria for comparing the models, and for selecting the most appropriate model for each domain. Thereafter, we compile the results from the models, for each of the domains, to investigate the conclusions that we can draw about steering in online settings.

### 6.1 Model comparison

For all models, we report two measures of performance: The evidence lower bound (the ELBO), which is the lower-bound on the marginal log-likelihood of the data under the model assumptions [11,17,27]; and the mean square error (MSE) of the model for reconstructing the original number of actions for each user. Parameter estimation is done in Pytorch and Adam is used to maximise the ELBO with an initial learning rate of 0.01 [16]. The learning rate was decreased with an exponential decay. We set the dimensionality of the latent space to $m := 10$.

We first report the results for the badge studies in Table 4. All models are trained on 60% of the data, with 20% of the data left for a validation set for model selection and 20% of the data is held out for a test set. Table 4 compares the performance of the models on the same test set. The results from Table 4 show that Model 2 outperforms the other models

[8] https://meta.stackexchange.com/questions/201728.

**Table 4** ELBO and MSE on held out data for badge study

| Badge | Model 0 | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|
| | ELBO | MSE | ELBO | MSE | ELBO | MSE | ELBO | MSE |
| Electorate | −256.94 | 2881 | −254.6 | 2855 | −235 | 2629 | −239.0 | 2717 |
| Civic Duty | −137.3 | 798 | −137.1 | 794 | −133.6 | 761 | −132.7 | 754 |
| Copy Editor | −392.2 | 10,122 | −409.1 | 10,003 | −385.2 | 9951 | −408.9 | 10,609 |
| Strunk & White | −120.0 | 669 | −119.4 | 655.1 | −118.7 | 654 | −119.1 | 651 |

**Table 5** ELBO and MSE on held out data for reputation study

| Threshold | Model 0 | | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|---|---|
| | ELBO | MSE | ELBO | MSE | ELBO | MSE | ELBO | MSE |
| 1K | −20.3 | 90 | −20.2 | 88 | −19.9 | 87 | −20.0 | 87 |
| 2K | −19.4 | 53 | −19.4 | 52 | −19.2 | 50 | −19.0 | 50 |
| 20K | −63.1 | 315 | −57.4 | 263 | −56.1 | 260 | −58.6 | 268 |
| 25K | −63.2 | 417 | −71.4 | 511 | −62.6 | 435 | −63.4 | 455 |

achieving a higher bound on the marginal log-likelihood (ELBO) and a lower mean-squared reconstruction error on unseen data (MSE) on all instances except the Civic Duty (where it is still near optimal) and on the MSE metric for the Strunk and White badge. Models 1, 2 and 3 all outperform Model 0, suggesting that the inclusion of the steering parameter $\beta$ does increase the probability of the activity data. Similarly, Models 2 and 3 outperform Model 1 which suggests that the steering strength parameter, $S_u$, is a useful way to segment the population of users. However, the additional complexity of Model 3 does not appear to help the model in better describing the data.

Table 5 presents the results for the reputation thresholds study. We use the same 60%, 20% and 20% splits for the train, validation and test sets, respectively. Due to the large data sizes for the $1K$ and $2K$ thresholds, we limit the data to a maximum of $10K$ users for each of the splits. Similar to the badge study, we report the ELBO and MSE on the held out test set. We also extend Model 2 to allow for an additional response to the reward that might be present in the data: As a user unlocks a privilege, she might choose to interact more on the website to explore the newly available features [9]. Thus, this model has an additional steering parameter, $\beta^3$, that is restricted to be **0** before the threshold is reached and non-negative after the threshold is reached. The other models remain the same as those used in the badge study.

In general, the Models 2 and 3 do outperform Models 0 and 1; however, the differences in their performance are less pronounced than that observed in the badge study. This leads us to the same conclusion as above that the steering parameter $S_u$ plays an important role in segmenting the behaviour of the users, but we note that the simpler models still captures the interaction dynamics well which suggests a more homogeneous set of reactions to the threshold. We choose Model 2 as the simplest model that describes the data for these three thresholds (it is optimal for the $1K$, $2K$ and $25K$ thresholds and it is near optimal for the $2K$). In all cases, it is important to note how well Model 0 performs, suggesting that many users do not in fact deviate from their preferred distributions for interaction and the null hypothesis (that steering does not occur) is a broadly good hypothesis for these reputation point threshold domains.

## 6.2 Analysis of steering

This section studies separately the steering effect that is inferred by Model 2 on the Electorate badge population (Sect. 6.2.1) and the effect inferred by Model 2 on the $1K$ reputation threshold population (Sect. 6.2.2). Although we study in detail only the Electorate badge in particular, the conclusions that are reached for the other badges are similar and thus we omit them for clarity; replicated plots for these domains can be found in Appendix B. Similarly, our focus below is on the $1K$ threshold for the reputation study. There are some subtleties regarding the behaviour of the SO users when they cross the different thresholds; most notably, the user behaviour around the higher thresholds appears to be different to that when they cross the lower thresholds. When discussing these results, we note when the activity around a specific threshold departs from the general trend that we observe. As with the badge study, the replicated plots on the other datasets are available in Appendix c.

### 6.2.1 Analysis of steering towards badges

We analyse the inferred parameters from Model 2 on the Electorate dataset to make conclusions about how people steer towards badges. Model 2 allows for four different types of users:

Type 1 (Non-Steerers): Users who do not deviate from their preferred distribution. In this case $S_u = (0, 0)$ and there is no effect of $\beta$ on their activity.

Type 2 (Strong-and-Steady): Users who experience the full adherence to $\beta_\lambda$ on their activity parameter $\lambda_u$ but do not change how often they interact on the platform (e.g. in a given day, they will complete more work but they do not work on more days). In this case, $S_u = (0, 1)$.

Type 3 (Dropouts): Users who appear to work on more days before the badge has been achieved and on fewer days after the badge has been achieved, thereby experiencing the effect of $\beta_\alpha$. They do not appear to change the number of actions that they will perform on a given day. In this case, $S_u = (1, 0)$.

Type 4 (Strong-Steerers): Users who adhere to the full steering effect described by $\beta = (\beta_\alpha, \beta_\lambda)$, both on how often they act on the platform and on how many actions they are likely to perform on any given day. In this case, $S_u = (1, 1)$.

Figure 4 presents the inferred assignment of users to the four user types (when considering the entire dataset). We can clearly see that the most common assignment type is Type 1 (Non-Steerers) making up 63.2% (3602 users) of the user base. The next most common type is the Strong-and-Steady group (19.8%; 1130 users) followed by the Strong-Steerers (13.5%; 772 users) and finally the Dropouts (3.5%; 197 users). A key finding is that the largest group that is inferred in the data does not appear to respond to the badge incentives in a way that has been predicted by previous works [1,24,28]. We highlight the fact that this Non-Steerer population is twice as large as the Strong-Steerer and Strong-and-Steady groups together! While these "steering groups" form a smaller population of users, it is the highly engaged interactions from these users that drive the aggregated trends that we notice in Fig. 2a.

[htb]

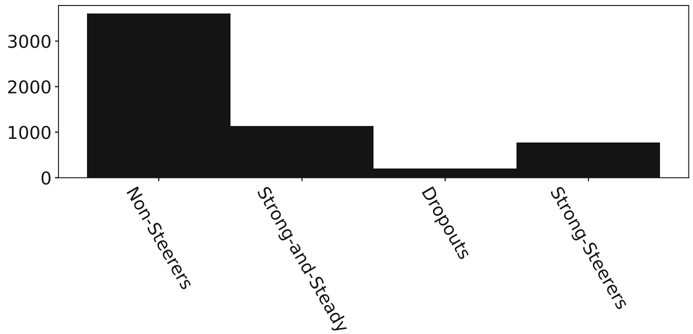


**Fig. 4** Cluster assignments (as inferred by $S_u$ from Model 2) for the users who achieved the Electorate badge

We demonstrate the markedly different behaviour of the users from each group by presenting samples from their interaction data, along with the models reconstruction of their activity. Figure 5 shows 10 random samples from these the who achieved the Electorate badge for each of the 4 user types. The plots show the true count of actions as a function of time alongside the expected number of actions under the assumptions of Model 2. The black vertical line, on day 0, corresponds to the day that the user achieved the Electorate badge. The left most column of Fig. 5 presents samples from the Type 1 (Non-Steerer) population. The counts of actions appear to show no change around day 0; these users appear not to change their behaviour in the presence of the badge. This is in stark contrast to all the other columns where there does appear to be a change around day 0. On the right hand column, we present samples from the Type 4 (Strong-Steerer) population of user. It is important to note the high number of actions (both expected and true) before day 0 when the badge was achieved. After day 0, both the true and expected numbers of actions drops dramatically. The center left column of Fig. 5 presents samples from the from the Type 2 (Strong-and-Steady) population. These users appear to increase the number of actions that they perform on a day leading into the badge achievement. They continue to work even after the badge has been achieved but at a reduced rate. This suggests that they have other reasons, than merely the badge, for contributing to SO. Finally, the center right column of the plot shows samples from the Type 3 (Dropouts). These users appear to hold a steady (and low) rate of interaction leading into the badge achievement followed by a decrease in their rate of activity after the badge is achieved.

Figure 6 shows the effect of steering on users, plotting $\beta$ as a function of time. The magnitude of the values of $\beta$ indicates direct changes to the probability that the user is active, as well as expected changes in the number of actions on a given day. In accordance with related work, and the predictions of the goal gradient hypothesis, users increase their rate of activity as they approach the day upon which they achieved the badge [1,6,24].

A novel insight from our model is that the $\beta_\alpha$ parameter, affecting both the Strong-Steerer and the Dropout groups, decreases well below 0 after the badge has been achieved. That is, users may decrease their activity well beyond their preferred distribution after they have achieved the badge. This result suggests that for some of the users who are steered strongly, they may stop contributing altogether once the badge has been achieved. This would align with a utility theoretic model of the behaviour where all the utility of the badge is conveyed upon receipt of the badge and thus there is no reason to continue to contribute [13]. This does

[hbtp]

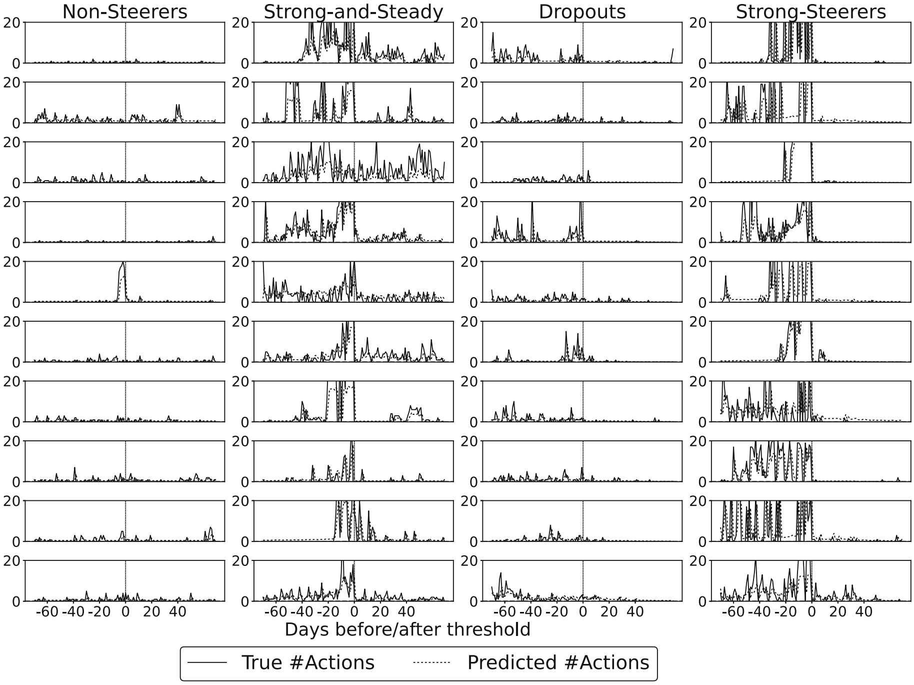
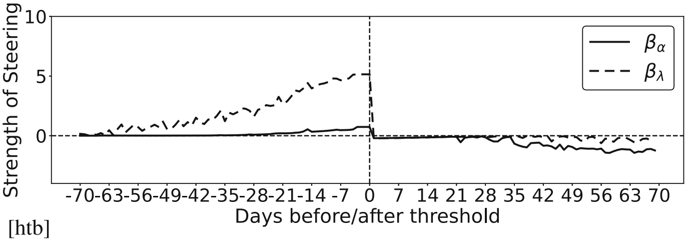


**Fig. 5** 10 samples of users' interaction data, with the corresponding model reconstructions, for each type of user as inferred by Model 2. Left column is the Non-Steerer group who appear to show no change in their behaviour around the badge achievement. Center-left is the Strong-and-Steady group who increase the number of actions they perform in a given day before achieving the badge. These users mainly continue to interact even after the badge has been achieved. Center-right presents samples from the Dropout users who appear to decrease their activity after achieving the badge. The right column presents the Strong-Steerer population who increase their rate of activity strongly before achieving the badge but decrease their activity rate to near 0 after the badge is achieved



[htb]

**Fig. 6** Plot of the inferred magnitude of $\beta$: the expected deviation from a user's preferred distribution $P_u$ under the assumptions of Model 2

not hold for all of the users as evidenced by the comparatively large size of the Strong-and-Steady population.

Figure 7 presents the mean number of interactions per user as a function of the number of days until/after the badge is achieved. The four lines correspond to the four groups that are inferred by Model 2. The mean interaction rates of these groups show the vastly different behaviours that are described above. In particular, we make the comparison of this plot to

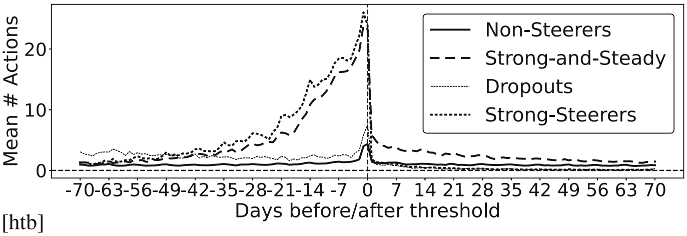


**Fig. 7** Mean number of actions per day for users who are classified by their steering parameters ($S_u$). The thin dotted line for the Dropout user group indicates that this group consists of less than 5% of the user base

that in Fig. 2a. We can see that the steering behaviour that is evident in Fig. 2a is actually mainly driven by the Strong-Steer and the Strong-and-Steady groups (together accounting for 33.36% of the population). Notice that the mean interaction count from the Strong-Steerer and Dropout groups drops passed the other groups to close to 0 after they achieve the badge. Of interest is the Strong-and-Steady group (13.6%) who act exactly as Anderson et al. [1] describes in that they return to a baseline level of work and continue to interact after the badge has been achieved. The thin dotted line for the Dropout user group is used to indicate that this group consists of less than 5% of the user base.

The Non-Steered population (63.2%) shows no change in their interaction rates before or after the receipt of the badge. There is a distinct uptick in the mean number of question-vote actions on the day before and on the day of the badge achievement (Fig. 7, blue line). It is possible that this “bump” might mistakenly be seen as the response of the users to the badge incentive. In fact, this bump is an artifact of the analysis technique which centers trajectories around a threshold that is crossed by the cumulative sum of the trajectory entries (see Sect. 7 and Appendix A for a discussion and proof of this claim).

### 6.2.2 Analysis of steering towards reputation points

In studying the response of the users to the reputation thresholds, we use the same grouping as that introduced above for the analysis of the Electorate badge. Figure 8 shows that the Non-Steerer population is again the dominant group that is inferred in this reputation threshold dataset. These users account for approximately 96.0% (68,941 users) of the user base whereas the Strong-and-Steady, the Dropouts and the Strong-Steerers only account for 1.6% (1146), 0.04% (34) and 2.3% (1674), respectively. The inferred fraction of Non-Steerers for the reputation thresholds is therefore greater than what is inferred for the badges thresholds. This holds for all the reputation thresholds and badges that we study in Appendices B and c.

Figure 9 shows the mean plot of activity for the groups, as inferred by the $S_u$ variable. The Non-Steering group is striking in that it shows the same low activity rates as those observed in Sect. 6.2.1 but for an even larger fraction of the population. The general trend that we observed in Fig. 3a are driven by the $< 5\%$ of the population who appear to respond to the badge. The Strong-and-Steady group shows the steering effect by a rapid increase in actions into the goal achievement, followed by a return to their base level of interaction. The thin dotted lines in the plot emphasise that each of these groups consist of less than 5% of the users who achieved the $1K$ threshold.

The Strong-Steerer group that was inferred for the $25K$ threshold consisted of 5.3% of the population with the Strong-and-Steady accounting for 4.0% (Figs. 28, 29 where the line for the Strong-Steering group is dark to reflect this). The higher portion of steerers for this

[htb]

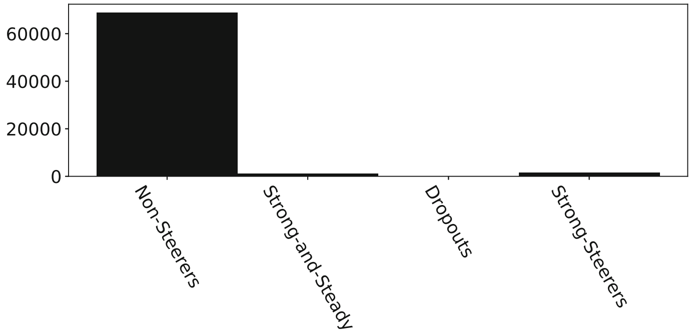


**Fig. 8** Cluster assignments, as inferred by $S_u$ from Model 2, for the users who achieved the 1K reputation point threshold

[htb]

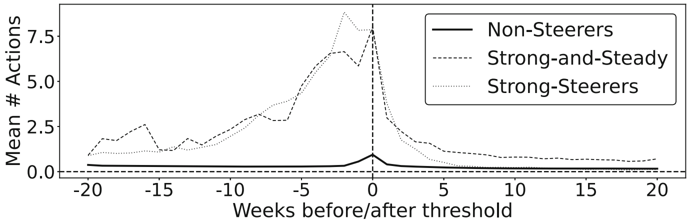


**Fig. 9** Mean number of actions per day for users who are classified by their steering parameters ($S_u$) for the users who passed the $1K$ reputation threshold

population could be due to the lack of further thresholds / privileges but we also note that the sample size for this population is the smallest and thus it could be due to the small sample who achieved this threshold.

### 6.3 Limitations of empirical study

The empirical study of steering that is presented in this section has a number of limitations which we list here. We only studying 4 of the threshold badges, 4 reputation thresholds and the study is limited to studying user behaviour on one platform: SO.

There are alternative types on badges that are present on SO. For example, the Famous Question badge[9] is awarded to a question that gets 10,000 views. It is not clear how users can “work towards” a “qualitative” badge of this nature and thus our study does not extend to badges of this type.

Secondly, we have focused our study on only 4 out of the total 26 privilege thresholds that SO defines. Our overarching conclusion is that steering is a rare phenomena in these settings but it is possible that there is a threshold where users have exhibited a greater steering effect than what we observed. Our choice of $threshold \in \{1K, 2K, 20K, 25K\}$ was motivated by the fact that the $1K$, $20K$ and $25K$ thresholds are three out of the five “milestone” thresholds on SO. Moreover, the $2K$ threshold provides a very well-defined privilege that may have resulted in a change in behaviour (as noted in Sect. 5).

[9] https://stackoverflow.com/help/badges/28/famous-question.

A final limitation is that the study was conducted only on SO data. While the goal gradient effect has been documented in many different domains domains Hull [12], Kivetz et al. [20], and steering has even been noticed on other question and answer platforms [6,24,28], our results are limited to the behaviour of users on the SO platform.

## 7 Proving the existence of phantom steering

The population of non-steerers in Figs. 7 and 9 display a sharp uptick in the mean of their action counts on the day before and on the day of the badge achievement. We prove that such a bump arises as an artifact of centering the data on day 0, and it is therefore expected to arise even in the absence of a steering effect. We show this "phantom steering" bump occurs in the setting of Model 0 (Fig. 1) where daily action counts are independent draws from some unchanging latent distribution. Our proof (and the intuition arising from it) suggests that a similar bump arises in the presence of steering as well. It is possible that this bump may have served to inflate previous conclusions about how users change their behaviour when working to achieve badges [1,24,28].

For users acting under Model 0 we present Theorem 7.1, which implies that for sufficiently large badge thresholds the expected number of actions on day 0 (the day of badge achievement) is greater than the expected number of actions on any other day.

We introduce this theorem via the following intuitive example: Suppose that the badge threshold $N$ is chosen randomly from some large range $N \in [m, M]$ of possible action counts. Let $S_n$ be the cumulative number of actions from a user up to (and including) day $n$. As long as the user continues to act on the platform, $S_n$ will eventually traverse the interval $[m, M]$. Moreover, as the count of actions on any day $n$ is a random variable (drawn from the user's preferred distribution), $S_n$ is more likely to cross the threshold $N$ on a day on which the user makes relatively more contributions. This claim holds even when actions are drawn under the no-steering assumptions of Model 0 which assumes that users' action counts on each day are independent draws from their preferred distribution $P_u$ (which is not influenced by steering).

We formalise this intuition in Theorem 7.1, the proof of which appears in Appendix A. Recall that the random variable $A_u^0$ describes the number of actions that user $u$ performs on the day that they receive the badge. Denote the number of actions required to achieve the badge by $N$, and let $A_{u,N}^0$ denote this random variable when the badge threshold is $N$ actions and user $u$ acts according to Model 0.

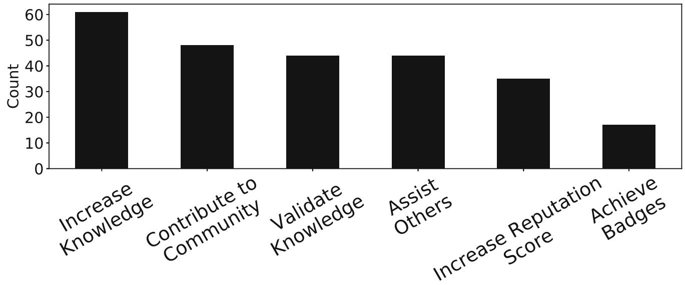
**Theorem 7.1** *If $P_u$ is bounded then:*

$$\lim_{N\to\infty} \mathbb{E}[A_{u,N}^0] = \mathbb{E}[P_u] + \frac{Var[P_u]}{\mathbb{E}[P_u]}.$$

This expected bump size holds in the limit as the badge threshold becomes large with respect to the mean of $P_u$. For fixed $P_u$ the convergence to this limit is exponential in the threshold.

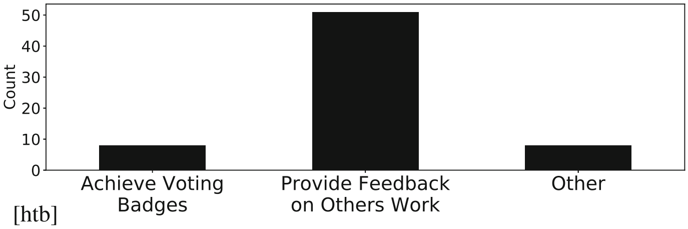
## 8 User survey

As an additional form of validation for the analytical results that are presented in this paper, we hosted a survey that recruited participants from SO to answer questions relating to their motivations for contributing to the website. A clickable advertisement was placed on SO and



[htb]

**Fig. 10** Counts of responses to the reasons for contributing to Stack Overflow



[htb]

**Fig. 11** Counts of responses to the reasons for voting in Stack Overflow

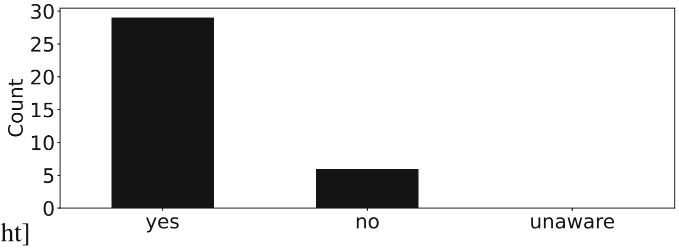
willing participants were directed to the survey. We paid each survey participant $10 in an Amazon gift voucher for completing the survey. In total, we received 86 responses from the community. We rejected 16 of these responses as the account IDs that were associated with these users did not exist or the users had not contributed to SO, making them not part of the target population. This left 70 valid survey responses.

Figure 10 summarises the responses to the question: “What are your reasons for participating in SO?” The majority of users claimed to have personal and/or altruistic reasons for contribution to the website with 87.1% claiming to contribute to increase their own knowledge (and 68.6% claiming to want to “contribute to the community”). In contrast to this, only 24.2% of the users selected the reason to “achieve badges”. 50% of users claimed that they had a goal of increasing their reputation score.

We also asked the users specifically about their voting contributions: “What motivates you to vote on other people’s posts?” The responses to this question are summarised in Fig. 11. Participants could select any combination of three different reasons for voting: badge acquisition (“I wished to achieve one of the voting badges, e.g. Supporter, Critic, Suffrage, Vox Populi, Civic Duty or Electorate”); altruism (“I think it is important to provide feedback about other’s work”), or “other”. Only 12.9% of participants who engage in voting actions reported badge acquisition is a motivating factor for their work. (Eight of the participants in the study claimed to not engage in voting actions and were not counted.)

Together these results present further evidence to corroborate the model predictions that only a minority of the SO participants are indeed steered by badges.

A surprising result, and one that stands in contradiction to the computational results presented in Sect. 6 is shown in Fig. 12. Participants were asked if “Privileges that are associated with a high reputation score motivate [them] to achieve a higher score?” The overwhelming response from the surveyed population was that these privileges did motivate



[ht]

**Fig. 12** Surveyed users' answers to the question: "Do the privileges that are associated with a high reputation score motivate you to achieve a higher score?"

the users, however, our results from Sect. 6 suggest that the steering hypothesis is weaker in this setting than in the badge setting (where the reward is more explicit).

We note the limitation of possible sample bias in the self-reported survey. A clickable advertisement was placed on the SO website and from there users opted-in to completing the survey. It is possible that the users who choose to complete such a survey have a biased perspective towards the rewards on SO. These biases would then show in our results. Moreover, we only had 70 users complete the survey and thus this represents a very small sample from the SO user base.

## 9 Conclusion and future work

We have presented a novel probabilistic model that describes how users interact on the SO platform and in particular how these users respond to badge incentives and to the reputation thresholds that unlock new privileges on the website. We demonstrated how this model can be fit to the data that is provided by SO and we investigated the distribution that is learnt over the latent space that describes the "steering effect".

Our results provide a more informed understanding of how users respond to badges in online communities. First, that some users do exhibit steering supports the claims made by previous work. These users comprise approximately 30% of the users for the badge studies and approximately 5% of the users for the reputation threshold studies. The users in this group significantly increase their levels of activity leading into the day when they achieve the goal. Some of the users, the "Strong-and-Steady" group, continue to interact at a base rate well after achieving the goal. This behaviour is well documented by previous work [1,22,24,28]. However, other users, the "Strong-Steerers" and the "Dropouts", actually decrease their activity rates, well below any previous level of activity, once the goal has been achieved. It is possible that assigning additional badges, with thresholds beyond those already in place in SO will continue to motivate such users.

Second, we identify the presence of a large population of users, approximately 60% of the population for the badge study, who do not exhibit steering. In the case of the reputation point thresholds, our results suggest that approximately 90% of the population does not exhibit steering. These "Non-Steerer" users do not appear to change the rate of their activity for the period under study (20 weeks for the badge studies and 40 weeks for the reputation points). Rather, they continue to act with the same rate well after the goal has been achieved. This suggests that these users have reasons for performing actions on SO which do not include specific receipt of the badge or privelege reward.

Third, any analysis of user behaviour around a goal must take into account the presence of the phantom steering bump which has not previously been acknowledged in the context of badges. This statistical artifact is model independent and may lead to inflated conclusions about the effect of badges on users' behaviour.

Future work will extend the models of Sect. 3 to study the feedback mechanisms on Q&A websites such as SO. While our empirical results suggest limited effect of the reputation privileges, our survey results suggest that the reputation points and the user generated feedback that drives this system, remains an important factor in motivating further contributions from the community. We believe that this relationship might depend on a tight feedback loop from action to response (vote or accept) and back to action. For example, a user who answers many questions, and receives the social validation from many "upvotes" and "accepts" (leading to reputation points), might experience an increased drive to continue interacting on the website. From examining SO comments and discussions, we also believe there is evidence that the misuse of votes could actually discourage participation. An example is when a user comments on the usefulness of an answer but fails to upvote and/or accept the answer. In some cases, there is evidence that the answerer feels demotivated by the lack of validation. The model that is presented in this paper can form the foundations for these works in that the $\beta$ parameters (i.e. the generic response to rewards) can be adapted to rather model this point process style of feedback data.

Moreover, we plan to investigate other aspects of feedback dynamics that are present on the SO platform. We believe that alternative theories from behavioural sciences, such as the sunk-cost effect, may apply to the users' behaviours and we plan to extend the models presented in this paper to consider such cases. These extensions will allow more detailed inference into who is motivated by the current feedback mechanisms and it will provide insight into how the feedback influences the behaviour of the users on such platforms.

**Acknowledgements** The authors would like to thank Stack Overflow for providing us with the voting data that allowed for this investigation. Moreover, Stack Overflow supplied advertisement space on its website such that a link to our survey was visible to their userbase. Hoernle is funded by a Commonwealth Scholarship. Kehne and Procaccia were partially supported by the National Science Foundation under grants CCF-2007080, IIS-2024287 and CCF-1733556; and by the Office of Naval Research under Grant N00014-20-1-2488.

## Appendix A: Omitted proofs

Here we present the proof of Theorem 7.1. Let $X$ be a nonnegative, bounded, and integer-valued random variable. Let $\{X_m\}_{m\in\mathbb{N}}$ be independent random variables which are distributed identically to $X$. We will be concerned with the partial sums $S_n = \sum_{m=1}^{n} X_m$. Let $Y_N$ denote the random variable which is the copy $X^m$ that brings $S_n$ across the threshold $N$; that is, for which $S_{m-1} < N$ and $S_m \geq N$.

**Theorem A.1** *If $X$ is nonnegative, integer-valued, and bounded then*

$$\lim_{N\to\infty} \mathbb{E}[Y_N] = \mathbb{E}[X] + \frac{Var[X]}{\mathbb{E}[X]}$$

More generally, we also consider the case when the $X$ are drawn from distributions $X^1, \ldots, X^\tau, \ldots, X^T$ repeatedly in turn. Then the partial sums are $S_n = \sum_{m=1}^{n} X^{m \mod T}$, where all copies of $X^\tau$ are independent. Let $\xi_\tau$ denote the event that $Y_N$ is drawn from distribution $X^\tau$, and let $Z = \sum_{\tau=1}^{D} X^\tau$. For this setting we have the following theorem:

**Theorem A.2** *If each of the distributions $X^\tau$ is finite, nonzero, nonnegative, and integer valued then*

$$\lim_{N\to\infty} \mathbb{E}[Y_N] = \frac{\sum_\tau \mathbb{E}[(X^\tau)^2]}{\mathbb{E}[Z]}.$$

Theorem A.1 follows directly from Theorem A.2 by taking the $X^\tau$ to be identically distributed. It therefore suffices to prove Theorem A.2.

We begin by showing that the likelihood of the sequence $\{S_n\}$ visiting any given number $N$ is asymptotically uniform. Let $p_m := \mathbb{E}\left[|\{n \in \mathbb{N} : S_n = m\}|\right]$, $g := \gcd(range(X))$ and observe that if $X > 0$ then $p_m = \Pr[m \in \{S_n\}]$. Also, if $m \notin g\mathbb{N}$ then clearly $p_m = 0$. For the $p_m$ for which $m \in g\mathbb{N}$, we have the following lemma:

**Lemma A.3** *If $X$ is nonzero, nonnegative, and bounded then*

$$\lim_{n\to\infty} p_{gn} = \frac{g}{\mu}$$

***Proof*** First, it suffices to assume that $g = 1$. This is because the integer-valued random variable $X' := X/g$ has mean $\mu/g$ and $\gcd(range(X')) = 1$, and proving the claim for $X'$ implies the claim for $X$. It also suffices to assume that $X > 0$. This is because the sequence $\{S_n\}_{n\in\mathbb{N}}$ remains at a specific value $m$ only so long as the independent draws are $X_n = 0$, after which it leaves $m$ forever. The expected number of steps that $\{S_n\}$ lingers at $m$ for is exactly $\frac{1}{1-\alpha}$, where $\alpha = \Pr[X = 0]$. Since $\mu > 0$ by assumption, we may prove the claim for $X'' := X|X > 0$. Then $\mu = \frac{\mu''}{1-\alpha}$ and

$$p_m = \mathbb{E}\left[|\{n \in \mathbb{N} : S_n = m\}|\right] = \frac{1}{1-\alpha} p''_m$$

Thus proving the claim for $X''$, proves the claim for $X$. Therefore, we may assume without loss of generality that $X > 0$ and that $\gcd(range(X)) = 1$.

Let $M := \max\{range(X)\}$ be the maximum value that $X$ obtains. Then the $p_m$ obey the recurrence

$$p_m = \sum_{j=1}^{M} p_{m-j} \Pr[X = j] \tag{6}$$

with the initial conditions $p_0 = 1$ and $p_m = 0$ for all $m < 0$. Because $X$ is bounded by $M$, we may break $\mathbb{N}$ up into "epochs" $\{1, \ldots, M\}, \{M + 1, \ldots, 2M\}, \ldots$, and then define $q_r^k := p_{kM+r}$ with $q^0 := (0, \ldots, 0, 1)^T$. For any $m = kM + r$ we can then iteratively expand the $p_{m-j}$ terms in Eq. 6 for which $m - j \geq kM$ until the expression for each $p_m$ depends only on the previous epoch, which gives an alternative recurrence of the form

$$p_{kM+r} = \sum_{s=1}^{M} \alpha_s^r \, p_{(k-1)M+s} \tag{7}$$

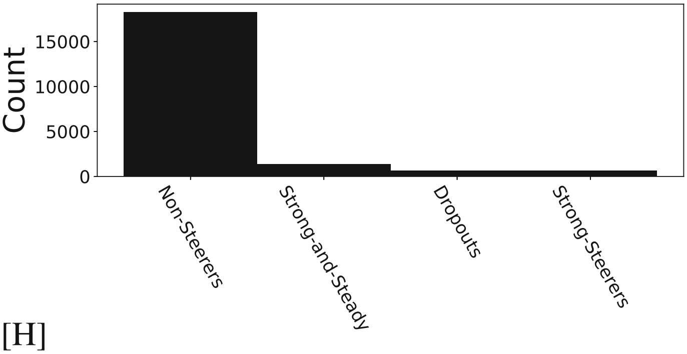
where $r, s \in [M]$ (and the initial conditions are the values of $p_s$ for $s \in [M]$). Note that these $\alpha_s^r$ do not depend on $k$. The recurrences in Eqs. 6 and 7 give $p_m$ as a convex combination of previous values, and so we may rewrite Eq. 7 as $q^k = A^k q^0$, where $A := \{\alpha_s^r\}_{r,s \in [M]}$ is a right stochastic square matrix. Furthermore it follows from the assumption $g = 1$ that $A$ is primitive. Therefore the Perron–Frobenius Theorem implies that $A^k$ converges exponentially quickly to a matrix of the form $\mathbf{1}\mathbf{u}^T$, where $\mathbf{1}$ and $\mathbf{u}^T$ are the unique right and left eigenvectors of $A$ corresponding to the eigenvalue $\lambda = 1$. This in turn implies that $q^k = A^k q^0$ converges to some uniform vector $(\gamma, \ldots, \gamma)$, and therefore that $\lim_{m \to \infty} p_m = \gamma$.

Finally we argue that $\gamma = 1/\mu$. We can show this by considering $C(N, J) := \mathbb{E}[|\{S_n\} \cap [N, J)|]$, the mean number of times that $\{S_n\}$ intersects some interval $[N, J)$. Since the $p_m$ converge, for fixed $J$ we may use linearity of expectation to choose $N$ large enough to guarantee that $C(N, J) \in J\gamma \pm \epsilon$ for any given $\epsilon > 0$. On the other hand, by considering the $\{S_n\}$ as "restarting" when they reach the epoch preceding $N$, we may use the central limit theorem to argue that $C(N, J) \in \frac{J}{\mu} \pm O(J^{2/3})$. Taking the limit as $J$ becomes large yields $\gamma = 1/\mu$. □

## Appendix B: Additional plots from badge study

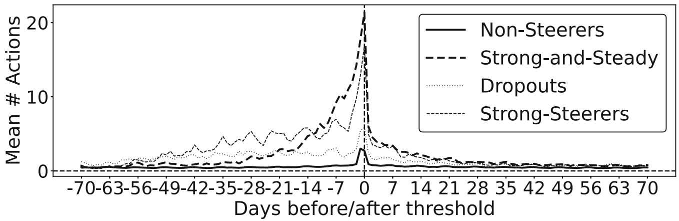
### B.1 Civic duty badge

See Figs. 13, 14 and 15.



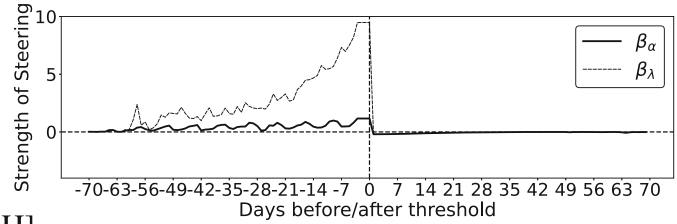
[H]

**Fig. 13** Cluster assignments (as inferred by $S_u$ from Model 2) for the users who achieved the Civic Duty badge



[H]

**Fig. 14** Mean number of actions per day for users who are classified by their steering parameters ($S_u$) for the users who achieved the Civic Duty badge

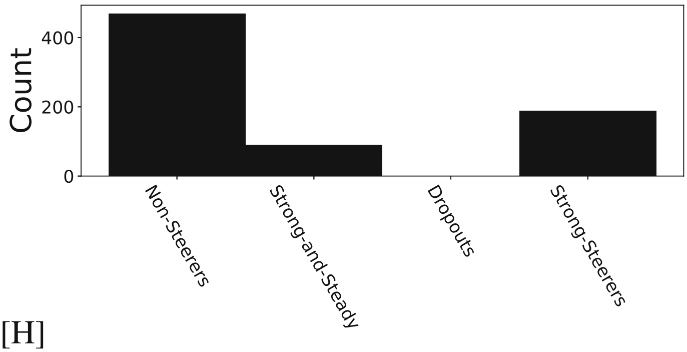


[H]

**Fig. 15** Plot of the inferred magnitude of $\beta$: the expected deviation from a user's preferred distribution $P_u$ under the assumptions of Model 2 for the users who achieved the Civic Duty badge

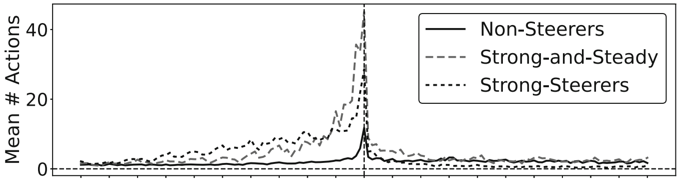
## B.2 Copy editor badge

See Figs. 16, 17 and 18.



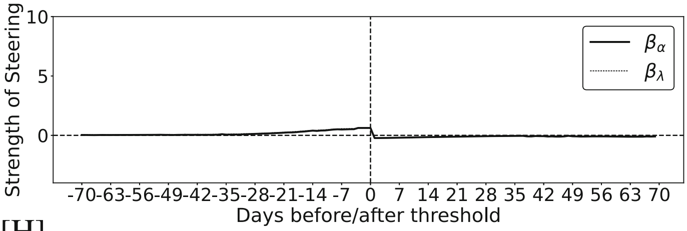
[H]

**Fig. 16** Cluster assignments (as inferred by $S_u$ from Model 2) for the users who achieved the Copy Editor badge



[H]

**Fig. 17** Mean number of actions per day for users who are classified by their steering parameters ($S_u$) for the users who achieved the Copy Editor badge

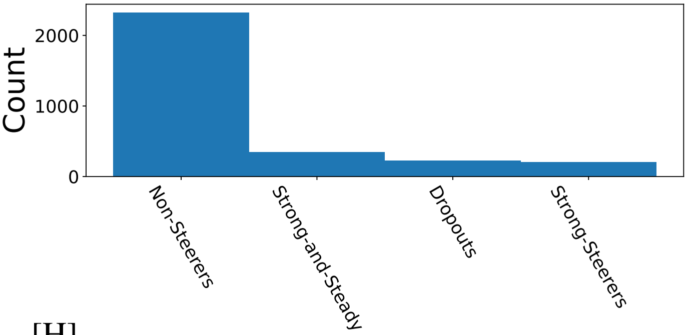


[H]

**Fig. 18** Plot of the inferred magnitude of $\beta$: the expected deviation from a user's preferred distribution $P_u$ under the assumptions of Model 2 for the users who achieved the Copy Editor badge

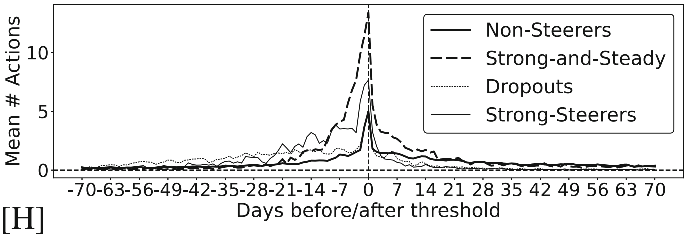
### B.3 Strunk and white badge

See Figs. 19, 20 and 21.



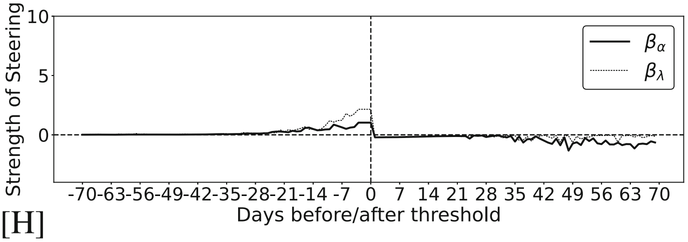
[H]

**Fig. 19** Cluster assignments (as inferred by $S_u$ from Model 2) for the users who achieved the Strunk & White badge



[H]

**Fig. 20** Mean number of actions per day for users who are classified by their steering parameters ($S_u$) for the users who achieved the Strunk and White badge



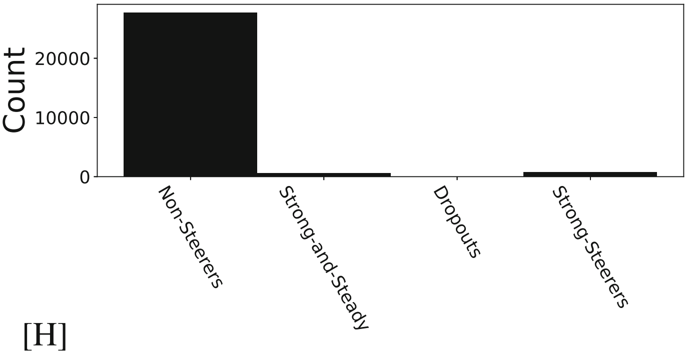
[H]

**Fig. 21** Plot of the inferred magnitude of $\beta$: the expected deviation from a user's preferred distribution $P_u$ under the assumptions of Model 2 for the users who achieved the Strunk and White badge

# Appendix C Additional plots from reputation threshold study

## C.1 Reputation threshold = 2*K*

See Figs. 22, 23 and 24.



[H]

**Fig. 22** Cluster assignments (as inferred by $S_u$ from Model 2) for the users who passed the $2K$ reputation threshold

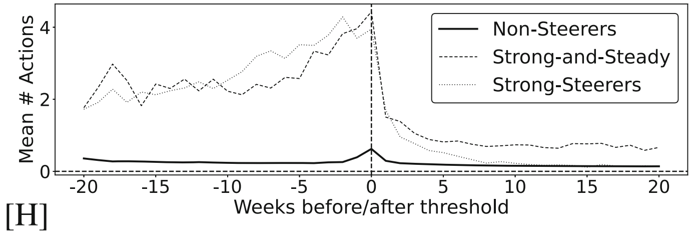


**Fig. 23** Mean number of actions per day for users who are classified by their steering parameters ($S_u$) for the users who passed the $2K$ reputation threshold

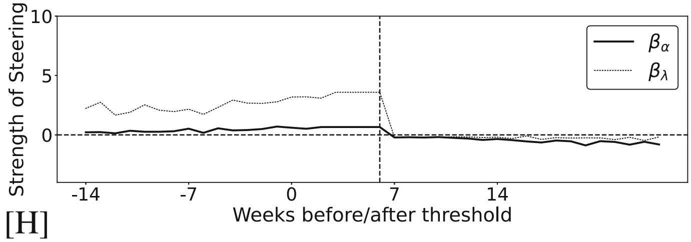


**Fig. 24** Plot of the inferred magnitude of $\beta$: the expected deviation from a user's preferred distribution $P_u$ under the assumptions of Model 2 for the users who passed the $2K$ reputation threshold

## C.2 Reputation threshold = 20*K*

See Figs. 25, 26 and 27.

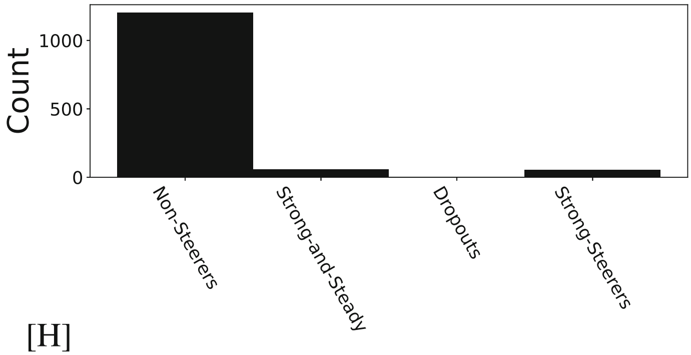


**Fig. 25** Cluster assignments (as inferred by $S_u$ from Model 2) for the users who passed the $20K$ reputation threshold

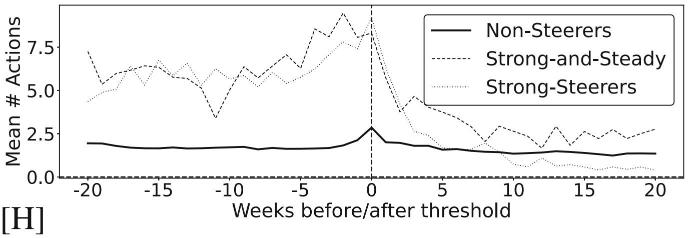


**Fig. 26** Mean number of actions per day for users who are classified by their steering parameters ($S_u$) for the users who passed the $20K$ reputation threshold

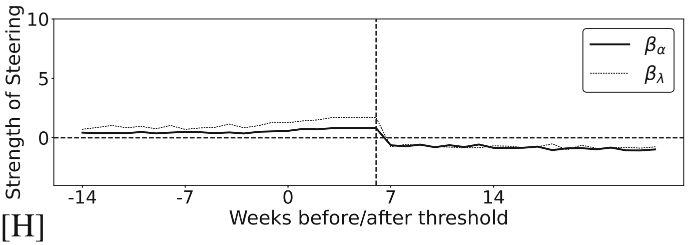


**Fig. 27** Plot of the inferred magnitude of $\beta$: the expected deviation from a user's preferred distribution $P_u$ under the assumptions of Model 2 for the users who passed the $20K$ reputation threshold

### C.3 Reputation threshold = 25*K*

See Figs. 28, 29 and 30.

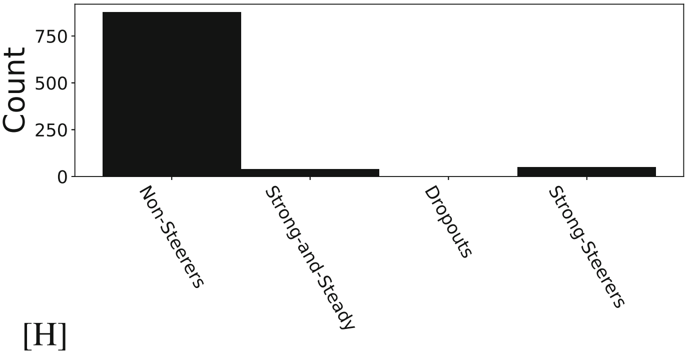


**Fig. 28** Cluster assignments (as inferred by $S_u$ from Model 2) for the users who passed the $25K$ reputation threshold

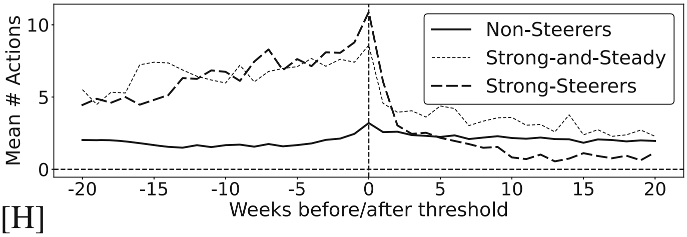


**Fig. 29** Mean number of actions per day for users who are classified by their steering parameters ($S_u$) for the users who passed the $25K$ reputation threshold

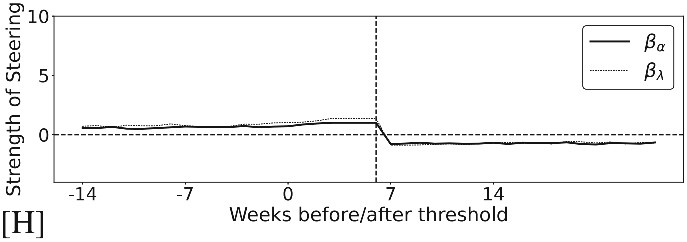


**Fig. 30** Plot of the inferred magnitude of $\beta$: the expected deviation from a user's preferred distribution $P_u$ under the assumptions of Model 2 for the users who passed the $25K$ reputation threshold

## References

1. 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. ACM, pp 95–106
2. 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. ACM, pp 687–698
3. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
4. Blei DM (2014) Build, compute, critique, repeat: data analysis with latent variable models. Annu Rev Stat Appl 1:203–232
5. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
6. Bornfeld B, Rafaeli S (2017) Gamifying with badges: a big data natural experiment on stack exchange. First Monday 22(6):1–17
7. Box GEP, Hunter WG (1962) A useful method for model-building. Technometrics 4(3):301–318
8. Burda Y, Grosse R, Salakhutdinov R (2015) Importance weighted autoencoders. arXiv preprint arXiv:1509.00519
9. Chou Y (2019) Actionable gamification: beyond points, badges, and leaderboards. Packt Publishing Ltd, Birmingham
10. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555
11. Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. J Mach Learn Res 14(1):1303–1347
12. Hull CL (1932) The goal-gradient hypothesis and maze learning. Psychol Rev 39(1):25
13. Immorlica N, Stoddard G, Syrgkanis V (2015) Social status and badge design. In: Proceedings of the 24th international conference on World Wide Web, pp 473–483
14. Ipeirotis PG, Gabrilovich E (2014) Quizz: targeted crowdsourcing with a billion (potential) users. In: Proceedings of the 23rd international conference on World Wide Web. ACM, pp 143–154
15. Jang E, Gu S, Poole B (2016) Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144
16. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
17. Kingma DP, Welling M (2013) Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114
18. Kingma DP, Rezende DJ, Mohamed S, Welling M (2014) Semi-supervised learning with deep generative models. arXiv preprint arXiv:1406.5298

19. Kingma DP, Salimans T, Jozefowicz R, Chen X, Sutskever I, Welling M (2016) Improved variational inference with inverse autoregressive flow. In: Advances in neural information processing systems, pp 4743–4751
20. Kivetz Ran, Urminsky Oleg, Zheng Yuhuang (2006) The goal-gradient hypothesis resurrected: purchase acceleration, illusionary goal progress, and customer retention. J Mark Res 43(1):39–58
21. Kusmierczyk T, Gomez-Rodriguez M (2018) On the causal effect of badges. In: Proceedings of the 2018 World Wide Web conference, pp 659–668
22. Li Z, Huang K-W, Cavusoglu H (2012) Quantifying the impact of badges on user engagement in online Q&A communities. In: International conference on information systems
23. Maddison CJ, Mnih A, Teh YW (2016) The concrete distribution: a continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712
24. Mutter T, Kundisch D (2014) Behavioral mechanisms prompted by badges: the goal-gradient hypothesis. In: International conference on information systems
25. Neal RM (1993) Probabilistic inference using Markov chain Monte Carlo methods. University of Toronto, Toronto
26. Ranganath R, Gerrish S, Blei D (2014) Black box variational inference. In: Artificial intelligence and statistics, pp 814–822
27. Rezende DJ, Mohamed S (2015) Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770
28. Yanovsky S, Hoernle N, Lev O, Gal K (2019) One size does not fit all: badge behavior in Q&A sites. In: Proceedings of the 27th ACM conference on user modeling, adaptation and personalization. ACM, pp 113–120
29. Zhang H, Wang S, Chen T-H, Hassan AE (2019) Reading answers on stack overflow: not enough! IEEE Trans Softw Eng 47:2520–2533

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Nicholas Hoernle** is a Ph.D. student in Computer Science at the University of Edinburgh, advised by Kobi Gal. His focus is on mechanism design in online collaborative systems and in designing and building probabilistic models that describe peoples' collaborative activities.

**Gregory Kehne** is a Ph.D. student in Computer Science at Harvard University, advised by Ariel Procaccia. His focus is on theoretical computer science, in particular on problems in computational social choice and online and approximation algorithms.

**Prof. Ariel D. Procaccia** is Gordon McKay Professor of Computer Science at Harvard University. He works on a broad and dynamic set of problems related to AI, algorithms, economics, and society. His distinctions include the Social Choice and Welfare Prize (2020), a Guggenheim Fellowship (2018), the IJCAI Computers and Thought Award (2015), and a Sloan Research Fellowship (2015). To make his research accessible to the public, he has co-founded several not-for-profit websites including Spliddit.org and Panelot.org, and he regularly contributes opinion pieces.

**Prof. Kobi Gal** is a faculty member of the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev, and a Reader at the School of Informatics at the University of Edinburgh. His work investigates representations and algorithms for making decisions in heterogeneous groups comprising both people and computational agents. He has worked on combining artificial intelligence algorithms with educational technology towards supporting students in their learning and teachers to understand how students learn. He has published widely in highly refereed venues on topics ranging from artificial intelligence to the learning and cognitive sciences.