Signals Matter: Understanding Popularity and Impact of Users on Stack Overflow

Arpit Merchant
IIIT-Delhi
arpitdm@gmail.com

Daksh Shah
IIIT-Delhi
daksh17336@iiitd.ac.in

Gurpreet Singh Bhatia
IIIT-Delhi
gurpreetb@iiitd.ac.in

Anurag Ghosh
IIIT-Hyderabad
anurag.ghosh@research.iiit.ac.in

Ponnurangam Kumaraguru*
IIIT-Delhi
pk@iiitd.ac.in

ABSTRACT

Stack Overflow, a Q&A site on programming, awards reputation points and badges (game elements) to users on performing various actions. Situating our work in Digital Signaling Theory, we investigate the role of these game elements in characterizing social qualities (specifically, popularity and impact) of its users. We operationalize these attributes using common metrics and apply statistical modeling to empirically quantify and validate the strength of these signals. Our results are based on a rich dataset of 3,831,147 users and their activities spanning nearly a decade since the site's inception in 2008. We present evidence that certain non-trivial badges, reputation scores and age of the user on the site positively correlate with popularity and impact. Further, we find that the presence of costly to earn and hard to observe signals qualitatively differentiates highly impactful users from highly popular users.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**; *Reputation systems.*

KEYWORDS

Crowdsourced Knowledge; Digital Signaling

ACM Reference Format:

Arpit Merchant, Daksh Shah, Gurpreet Singh Bhatia, Anurag Ghosh, and Ponnurangam Kumaraguru. 2019. Signals Matter: Understanding Popularity and Impact of Users on Stack Overflow. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308558.3313583

1 INTRODUCTION

Stack Overflow has evolved from a simple Q&A site to a massive social community where knowledge seekers and knowledge providers of all levels of expertise interact with each other to solve programming difficulties [2]. It has significantly impacted the way programmers learn, communicate and collaboratively build content repositories for future reference [6, 31, 34]. Due to its widespread use, it has become an integral part of the software development ecosystem and developers increasingly rely on it for their daily programming needs. Moreover, users on other platforms such as mailing lists, Github, etc. actively encourage their participants to refer back to posts on Stack Overflow for solutions [38].

This rise in the site's importance stems from four factors: (i) users can find multiple high quality answers for questions on nearly every programming language, tool, framework and software [28], (ii) if what is needed is not available, they can create a post themselves and receive answers extremely quickly [24], (iii) virtual rewards (reputation points and badges) incentivize users to contribute [14], and (iv) the rich interface enables them to display their expertise to potential recruiters [9]. These factors facilitate the transparent nature of the site. Each user has their own dedicated profile page that aggregates their contributions and achievements on the site. As a result, other users and recruiters can form impressions about their expertise of topics, their programming abilities, skills and experience [13]. In such a highly competitive environment, users that stand out are those that successfully acquire visible traces to attract attention [12]. One such significant way for users to stand out is by acquiring a large number of reputation points and badges.

Social status and reward system design. Virtual rewards act as symbols of social status, despite having no explicit value of their own. Some badges require users to expend costly effort and are therefore earned by few. These confer a higher status value since they distinguish members within the community. Others are easier to earn and act as motivations, and sources of learning. Badges serve various socio-psychological functions on crowdsourced platforms [4]. A prominent theme in literature has focused on the roles of badges in incentive structures [10, 19, 22, 27, 29]. Immorlica, et al. [21] show that the optimal design employs threshold badges where only users above a pre-defined number of contributions receive badges. Easley, et al. [16] take a game-theoretic approach to analyze the effectiveness of systems of such threshold badges.

Effects of virtual rewards. A separate line of research has analyzed the qualitative and quantitative effects of virtual rewards in diverse settings such as open-source software [29, 37] and knowledge repositories [38, 39]. Anderson, et al. [3] define a formal model that predicts how badges steer user behaviour. Mutter, et al. [26] provide empirical evidence that as users' proximity to goals defined by badges increases, so does the level of the users' contributions (goal-gradient hypothesis). First-time badges, awarded after a user takes a specific action for the first time, causally affect user behaviour and

*This work was done in part while Ponnurangam Kumaraguru was on sabbatical at IIIT Hyderabad.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313583

Total number of active users	3,831,147
Total number of questions	15,711,957
Total number of answers	24,492,236
Mean reputation of active users	111
Mean number of badges earned by active users	22
Mean year when users joined the site	2015

Table 1: General statistics of users' activities in our dataset.

also improve the functioning of the site itself [23]. The reputation points of users on Stack Overflow along with community activity dynamics are good predictors of the long-term value of questions and answers [2]. But, what attributes these virtual rewards can signal about users themselves is not yet well understood.

Research Questions. In this paper, we focus on finding important markers of user attributes since they are known to relate to dynamics of identity, crowdlearning, social benefits and societal acceptance [5, 36]. Specifically, we ask:
RQ1: According to Stack Overflow users, what social qualities (if any) do reputation scores and badges intend to signal?
RQ2: To what extent do these game elements actually signal or indicate the qualities that users expect them to?
The paper most closely related to ours is that of Trockman, et al [35]. They analyze various categories of badges such as Quality Assurance, Dependency Management, etc., in the `npm` ecosystem on Github as signals of *repository properties* such as dependency freshness, test suite quality and popularity. Some of these signals are subjective. Also, maintainers of the repositories can choose which badges they wish to display and which they do not. We consider Stack Overflow with a completely different and more complex system of reputation points and badges that it awards *to users* and is based on objective, pre-defined metrics [3].
We summarize our contributions below.

- We conduct a survey of Stack Overflow users and draw preliminary insights about how they view reputation points and badges as indicators of various social qualities.
- We perform empirical investigations on a large dataset of 3,831,147 users and the complete time-stamped history of their actions on Stack Overflow spanning a decade.
- Employing nonlinear regression models, we find that the presence of certain non-trivial badges correlates with higher popularity and impact. We also provide evidence that badges add more explanatory power compared to reputation scores.
- Statistical analyses of user activity show distinct differences in patterns of engagement between popular and impactful users.

Through these findings, we shed new light onto the role of virtual rewards in studying user qualities on crowdlearning platforms.

2 THEORETICAL FRAMEWORK

The widespread adoption of game elements on Stack Overflow invites a deeper examination of their effects on its users. Reputation scores are received for taking various positive actions whereas badges are awarded for "being especially helpful". We argue that given the variety of actions rewarded through reputation scores and badges, they are important signals of underlying qualities of users. We thus investigate their value from a signaling perspective.

Adverse Selection. Users on Stack Overflow possess different levels of information about various topics as well as other users on the platform. Users have a better understanding of their own expertise and limitations. They thus choose to participate selectively in order to maximize their benefits. Users however, tend to be uncertain about the preferences of heterogeneous audiences in terms of how they will respond to their actions. At the same time, the audience's qualitative assessment of users' abilities is based on limited information. Such a state where neither party has complete knowledge about the other is called information asymmetry [32]. This causes adverse selection, i.e. bias towards only particular kinds of actions [30]. For instance, most individuals prefer high returns and so they differentially choose low-hanging fruits, and broadly useful actions, while a few others may prefer more niche and challenging questions. As a result, participation is severely affected.

Digital Signaling. Signaling is a well-studied and popular solution to the problem of adverse selection [20, 32]. Signals are images, symbols and signs that allow users to communicate information and meaning with appropriate context. Signals that are costly to generate for the signaler and cognitively easy to process for the observer tend to be very reliable [11]. The design of sets of such assessment signals can specifically combat the inefficiencies arising due to information asymmetry [15]. The audience on Stack Overflow upvotes or downvotes posts to indicate that they approve or disapprove of them. This is a basic signal that is cheap to produce. Conversely, reputation scores and badges can help highlight deep technical qualities of a user since they require significant effort to achieve. This allows the user to potentially make better decisions in the future and the audience to gain more knowledge about him/her.

Gamification. Gamification is the use of game design elements in non-game contexts [14]. Badges on Stack Overflow are automatically earned by users based on their performance, unlike Github, where they are voluntarily displayed [7]. A single badge can holistically combine multiple qualitative actions whereas reputation points can be earned for every positive unit of action. This involves users in a social environment thereby motivating increased participation. In this paper, we consider game elements such as reputation scores and badges to be *digital signals* and investigate whether they are indicative of the performance and qualities of users.

3 DATA DESCRIPTION

Our experiments are conducted on a publicly available dataset containing all individual time-stamped actions of Stack Overflow users from the site's inception on July 31, 2008 to June 5, 2018 [33]. Table 1 describes a summary of the general statistics of our data.

Reputation. Reputation[1] scores are officially considered a "rough measurement of how much the community trusts you". Reputation is earned (or lost) when a user's question or answer is upvoted (or downvoted), when an answer is marked accepted by the user who originally asked the question, when bounties are received (or spent), or when suggested edits are accepted.

Badges. Badges are awarded in addition to reputation scores when the corresponding pre-defined set of actions and/or reactions are performed. They can be classified in two primary ways.

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

– *Class-wise:* Bronze class badges are the easiest to obtain, Silver class badges require additional effort and Gold class badges are the hardest to earn. For instance, *Popular Question* is a bronze badge awarded to users when they ask a question that receives at least 1,000 views. *Notable Question* (silver) and *Famous Question* (gold) badges are awarded when the question receives at least 2,500 and 10,000 views respectively.
– *Category-wise:* Categories include Question, Answer, Participation, Moderation, Documentation, Tag and Other. Each category incentivizes users to conduct different kinds of actions. A user can earn multiple category badges as long as the requirements are fulfilled each time. For instance, a user obtains a new Nice Question badge for every question with a score of 10 or more.

Currently, there are 91 different badges[2] available on the site.

Active Users. We define active users as those who have asked at least one question, or have written at least one answer. We only consider these participants for our experiments to reduce the noise introduced by non-active users. Some users create throwaway accounts to ask a question. Our results hold when considering users who created up to 10 posts (questions and answers combined).
We also use other information available such as the time-stamps when users joined the site, number of questions, answers and comments they made, etc. The complete list of fields in the dataset along with detailed descriptions is available here.

4 USER SURVEY

We conducted an online survey of Stack Overflow users to gauge their views about the game elements and the platform in general.

Survey Design. We extracted email addresses of 2,740 users who had voluntarily shared this information in the About Me section of their profile pages on Stack Overflow. We divided these users into three groups based on their reputation scores (low, medium and high). Then, we randomly selected 500 users from each group and sent personalized invitations to participate in the survey. We received a total of 56 responses. Our respondents have a mean of 10 years of experience with coding/programming.
The survey[3] focused on two themes namely, (a) inferences regarding what reputation scores and badges can say about users, and (b) perceived effects of these game elements on the community. We also requested participants to indicate names of specific badges they considered important along with free-text boxes for longer comments, if any. The survey was piloted first.

Survey Results. The general consensus is that Stack Overflow is a good site to get multiple high-quality answers to programming questions. Across the three groups combined, 87% of the respondents either Strongly Agree or Agree with the statement *"Stack Overflow, in general, is more trusted than other communities for programmers."* Respondents felt that reputation scores tend to convey engagement, experience, contributions, helpfulness and knowledge. They consider badges like Good Answer, Popular Question, Pundit, Necromancer and Populist to be important. However, some answers mentioned that they did not consider reputation scores and badges to be important at all. Sixty five percent of respondents agreed with the statement that reputation scores are indicative of helpfulness but only 51% somewhat agreed that they indicate knowledge.

Survey Insights. Users interpret the importance of badges differently, yet badges and reputation scores have a subconscious impact on their future actions [23]. Our survey responses support these results. But they also point towards the question of what user attributes they can signal and to what extent. We focus on the following two attributes:
– *Popularity:* Respondents with low reputation consider reputation scores to be better indicators of user popularity than badges; yet majority of respondents with high reputation only somewhat agree with this statement.
– *Impact:* Fifty one percent of respondents feel that badges are a good measurement of how helpful and knowledgeable a user is (while another 30% somewhat feel the same). A larger majority, 67%, find reputation scores to be good indicators of expertise.

Note, we do not conclude our survey respondents' views to be representative of the entire community. Rather, we use these insights to design hypotheses which we then test empirically on the massive dataset described in Section 3.

5 CHARACTERIZING EFFECTS OF SIGNALS

Our goal is to identify important signals of popularity and impact of users based on their behaviour and actions on Stack Overflow.

Hypotheses. Based on survey insights, we test the following:

H1: Reputation scores and Badges are positively correlated with popularity as well as impact of users.
H2: Reputation scores are better indicators of popularity as well as impact compared to Badges.

Operationalization. To operationalize the two attributes in question, we adopt measures proposed by users, moderators and administrators on Meta Stack Exchange (a sister site for discussions on the workings and policies of Stack Overflow).
– *Popularity Score:* A user's place in the social landscape of Stack Overflow is the result of how they are perceived and how well they are known. We define the perceived popularity of a user to be the total number of distinct views on their profile page.
– *Impact Score:* A user's reach on the site is the number of people who have benefited from the user's actions. We consider the impact score of a user, as defined on Meta Stack Exchange[4], as the sum total views on questions, and answers with non-zero scores that have either been accepted, or are in the top 3 answers, or have a score of at least 5, or have at least 20% of the vote count.

Data Preparation. We preprocess our data as follows; (1) Since the distributions of both popularity and impact scores are heavy-tailed, we z-score transform them to capture the relative variation across users. (2) We create three sets of features namely, control, reputation, and number of badges. The Control Model (CM) consists of features such as number of days since the user joined the site, number of questions asked, number of answers given, etc. The Reputation Model (RM) consists of all the control features and one additional feature, i.e. the reputation score of the user. And the Badges Model (BM) similarly consists of all the control features plus

[2] https://stackoverflow.com/help/badges
[3] Complete survey questionnaire as administered to participants available here.
[4] Definition of impact score of users - A discussion on Meta Stack Exchange.

Features	Control Model	Reputation Model	Badges Model
Age on the site	**0.319**	0.225	**0.191**
Number of questions	0.055	0.074	0.008
Number of answers	0.250	0.047	0.075
Number of upvotes	0.122	0.123	0.021
Number of downvotes	0.115	0.092	0.048
Reputation score		**0.313**	
Nice Answer Badges			0.062
Populist Badges			0.052
Enlightened Badges			0.029
Necromancer Badges			0.039
Good Answer Badges			0.031
	$R^2 = 0.911$	$R^2 = 0.939$	$R^2 = 0.957$

(a) Regression models for predicting Popularity of users.

Features	Control Model	Reputation Model	Badges Model
Age on the site	**0.321**	0.225	0.065
Number of questions	0.129	0.129	0.015
Number of answers	0.250	0.094	0.119
Number of upvotes	0.085	0.067	0.013
Number of downvotes	0.033	0.049	0.006
Reputation score		**0.394**	
Great Answer Badges			0.069
Revival Badges			0.055
Enlightened Badges			0.071
Necromancer Badges			**0.188**
Good Answer Badges			0.043
	$R^2 = 0.685$	$R^2 = 0.767$	$R^2 = 0.858$

(b) Regression models for predicting Impact of users.

Table 2: Summary of importances of the exogenous variables for the Control, Reputation and Badges models. Table 2a and Table 2b show the results for predicting the popularity and impact of users, respectively. In each case, we present scores of the five most important badges. The importance values are relative to other exogenous variables within the specific model only.

one additional feature per badge, i.e. number of each such badges earned by the user (for all 91 badges on the site). We conduct an ablation study to compare their performances.

Model Fitting. We propose a Gradient Tree Boosting Regression model[5] to analyze the fit of the endogenous variable (popularity or impact score) from the exogenous variables (feature sets). We set the maximum tree depth as 3, learning rate as 0.1 and the number of boosted trees to fit as 100. We divide the data into training and testing sets and average the results over 50 runs of the experiment. We validate the model using the R^2 metric. Since it denotes how well the model fits the data points, higher values are better.

We also compute the relative importance scores of the features in each model. This score estimates the improvement in the squared error risk due to each feature compared to that for a constant fit [18]. Specifically, it is the average total decrease in impurity of a node across all trees in the ensemble. Decrease in impurity is the number of times a feature is used to split a node divided by the number of samples that it splits. It thus indicates how useful the feature was in the construction of the boosted decision tree model.

5.1 Signals of Popularity

Results. Table 2a presents the relative feature importances for fitting popularity scores within CM, RM and BM. For brevity, we report only the top five badges ordered according to their importance scores. BM explains 95.7% of the variance, while RM and CM explain 93.9% and 91.1% of the variance respectively.

Analysis. We observe RM provides more explanatory power compared to CM with a small, but significant increase in R^2 scores. This improvement in the model fit is due to the reputation points feature which also has the highest importance score. This indicates that it is a good predictor of popularity. That is, users with high reputation points tend to attract other users to their profile pages.

[5]The code is available on Github at this url.

Notably, BM outperforms with RM and CM in terms of goodness-of-fit. This happens because badges aggregate various sets of actions thereby providing more information than just the reputation score. For instance, the most important badge feature is the number of Nice Answer badges. This badge is earned every time a user provides an answer that receives a score of 10 or more. More generally, we find that the five most important badges are all Answer Badges.

5.2 Signals of Impact

Results. Table 2b similarly presents the performances of the three models in fitting impact scores. CM, RM and BM models achieve R^2 scores of 68.5%, 76.7% and 85.8% respectively. Here too, we find that BM significantly outperforms the other two models.

Analysis. Once again, we find that reputation points are good predictors of impact. Yet, BM improves upon RM and CM because badges capture a more nuanced summary of the user's contribution. Reputation points increase not only due to upvotes on posts, but also on performing other actions such as useful edits, winning bounties, etc. This combines all positive actions into a single score thereby diluting it's effect. Contrasted with BM, consider the number of Necromancer badges earned by the user. A Necromancer[6] badge is awarded on posting an answer to a question at least 60 days after it has been asked and which receives a score of 5 or more. Two answers with the same score, but one written on the same day (say) and another written 60 days after the question was posted represent different value to the community. This is captured by the badge and not the reputation score. It is interesting to note, that once again each of the 5 most important badges are Answer badges.

As we can see in Table 2b, the importance score of the best feature is almost twice that of the next best feature in RM and BM both. But these two models exhibit largely different performance characteristics. Due to the smaller number of features, RM is extremely

[6]https://stackoverflow.com/help/badges/17/necromancer

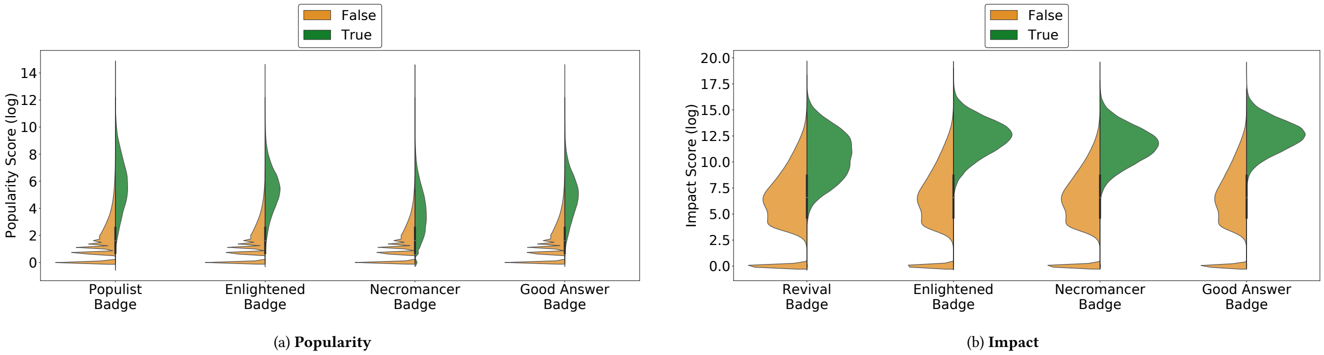


Figure 1: Distribution of popularity and impact scores of users (in log-scale) in the presence/absence of badges.

fast. BM is slower because it has more features and, in the context of Stack Overflow, they suffer from multi-collinearity. However, BM is more accurate because the ensemble of decision trees is able to separate users that have made helpful contributions on a variety of different metrics of helpfulness.

5.3 Discussion

Hypotheses. Our empirical results on the active Stack Overflow community mostly support our hypotheses and reveal interesting observations. High reputation corresponds with higher popularity and impact. Not all badges are good predictors; Documentation and Other badges show no correlation at all. However, certain Answer badges (such as, Necromancer, Enlightened and Good Answer) tend to be reliable signals. Figure 1 compares the distribution of users along their popularity and impact scores in the presence and absence of badges. This shows that users having these specific answer badges tend to be more popular and impactful than users that do not. The time since the user joined the site is an important signal of impact. And lastly, contrary to the impressions of users in our survey and the subsequent hypothesis (**H2**), badges are better predictors of popularity and impact compared to reputation.

Implications. Broadly useful and important questions can be asked by experienced and novice users alike. But writing impactful answers sometimes requires domain expertise. Different answers to the same question can be helpful to different users or to the same user at different times. Our results indicate that reputation scores seemingly fail to capture such nuances, whereas Answer badges appear adept at doing so. One potential explanation is that answers to some relatively easier questions with long-term value may have been posted during the initial years of the site [2]. Such answers yield a substantially high return on reputation to the original poster. Moreover, there exists evidence to suggest that reputation scores are easier to "farm" than badges through strategies such as writing answers in niche communities, or during off-peak hours [8].

Threats to Validity. We identify three primary threats to the validity of our approach. First, our metrics for computing popularity and impact scores are reductive. They are biased towards estimates of the number of views on profile pages and user posts obtained via internal site analytics. Second, we focus specifically on reward-based features and do not incorporate content-based features. Future work could examine linguistic attributes of posts that affect performance. Third, past evidence [17, 25] has shown that women have faced significant barriers to participating on the site. This suggests that game elements may be biased against some users. Thus, we advise caution in inferring broader interpretations of our results since we do not guarantee whether the positive links between badges and user attributes are causal or not.

6 DIFFERENTIATING POPULAR AND IMPACTFUL USERS

We now ask whether these game elements act as *differentiating signals* between popular and impactful users, and if so, why.

Thematic Representation of Users. Figure 2 depicts user distribution along two axes, popularity and impact. Time since the user joined the site is strongly associated with their standing in the community. As expected, a large concentration of users have low popularity and impact. Most new users along with a large fraction of the older users belong to this category. On the other hand, most highly popular and impactful users joined during the early years.

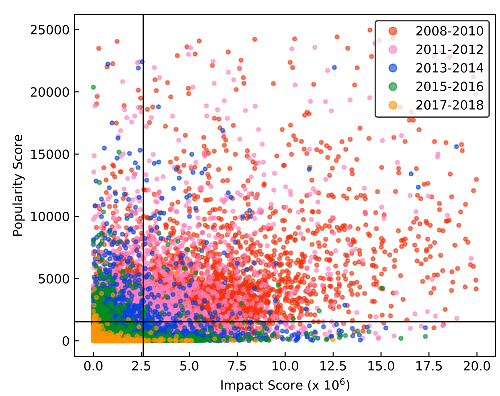
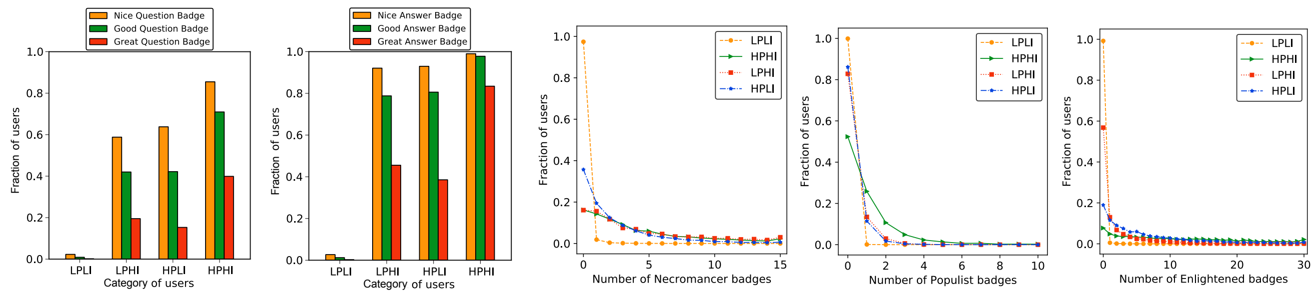


Figure 2: Distribution of users across popularity and impact scores based on the year in which they joined Stack Overflow. The horizontal and vertical black lines segment the population into the top 0.1%.

We segment users into four groups based on whether or not they belong to the top 0.1% of the community along the two social attributes: (a) high popularity, high impact (HPHI), (b) high popularity,



(a) **Fraction of users with badges for high scoring questions** (b) **Fraction of users with badges for high scoring answers** (c) **Distribution of number of Necromancer badges earned** (d) **Distribution of number of Populist badges earned** (e) **Distribution of number of Enlightened badges earned**

Figure 3: Relationship of LPLI, HPLI, LPHI and HPHI users to badges earned. Figures 3a and 3b study the presence of good quality question and answer badges among the different groups. Figures 3c, 3d and 3e depict the fraction of users in different communities that have earned multiple Necromancer, Populist and Enlightened badges.

low impact (HPLI), (c) low popularity, high impact (LPHI), (d) low popularity, low impact (LPLI). This segmentation is represented by the horizontal and vertical black lines in Figure 2. HPHI have mean popularity and impacts scores of 13,134 and 11,972,950 respectively, whereas LPLI have mean scores of 23 and 35,081 respectively. This shows the vast gulf between the two groups.

Figures 3a and 3b show the fraction of users belonging to each of the four categories HPHI, HPLI, LPHI and LPLI that have the particular badge. Interestingly, more LPHI, HPLI and HPHI users have badges for well-received answers (such as Nice Answer Badge) than they do for well-received questions (such as Nice Question Badge). Figures 3c, 3d and 3e display the distribution of the number of Necromancer, Populist and Enlightened badges earned by the four groups of users respectively. Consider the case of the Enlightened badge. We see that nearly 60% of LPHI users have zero Enlightened badges whereas only about 20% of HPLI users do not have that badge. We argue that there must be meaningful explanations that can be learned by comparing between these two groups.

Feature	HPLI	LPHI	t-statistic	Sig
Questions	54.65	42.02	-6.23	***
Answers	452.63	137.89	-39.97	***
Question Scores	233.15	286.61	4.38	**
Answer Scores	1190.07	679.83	-24.09	***
Reputation	16304.64	8672.31	-30.20	***
Necromancer Badges	2.47	6.32	29.6	***
Populist Badges	0.174	0.218	4.041	**
Great Answer Badges	0.682	0.887	7.84	***

Table 3: Differentiating between HPLI and LPHI users. ** = p < 0.01, *** = p < 0.001 represents statistical significance of Welch's t-statistic after Bonferroni correction ($p/14$).

We therefore examine HPLI and LPHI users and expect there to be differences in the way they contribute as well as reception to their contributions. Using Welch's t-test, we study the differences present between these two groups and present the features with the most significant differences between them in Table 3. We find that the number of questions and answers posted are significantly higher among HPLI, reflecting that they are more active. Conversely, the number of Necromancer, Populist and Great Answer Badges are higher for LPHI users. These badges appear to be signals that are costly to earn but not easily observable. Site design dictates that upvotes on answers return double the reputation points compared to upvotes on questions. Our findings show that LPHI users have a proportionally higher number of question and answer posts/scores. This implies that answers drive popularity, but it is questions that offer more influence. Further, some users link their SO accounts with other platforms such as LinkedIn, Github, etc. that may explain why they may be better known [1]. This is another potential source of divergence between high popularity and high impact.

7 CONCLUSION

The diverse range of actions and users, and massive quantity of content on Stack Overflow obfuscates the quality of information and efficiency of deliverables. It increases the transaction costs of participation. Game elements such as badges and reputation scores aim to provide incentives to balance these costs. But the design of these incentive structures has led to problems of adverse selection. In this paper, we present evidence that some of these game elements also act as reliable digital signals of social qualities such as popularity and impact. Our experiments reveal that certain non-trivial answer badges, high reputation scores and age of the user on the site indicate significant correlations. We also find differentiating characteristics that distinguish communities of popular and impactful users. We believe these insights offer guidance on combating inefficiencies arising out of bias towards specific actions. Our results encourage further exploration of the role of game elements as symbols of social status in socio-technical systems.

8 ACKNOWLEDGEMENTS

We thank Indira Sen, Shubham Singh, Shwetanshu Singh, and Srishti Aggarwal for useful discussions and the anonymous reviewers for detailed, helpful comments. We thank Language Technologies Research Centre (LTRC, IIIT-H) and Precog (IIIT-D) for their support. We are grateful to the respondents of our survey and to Stack Overflow for releasing their data publicly.

REFERENCES

[1] Kristen M Altenburger, Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, and Jim Hamilton. 2017. Are there gender differences in professional self-promotion? an empirical case study of linkedin profiles among recent mba graduates. In *Eleventh International AAAI Conference on Web and Social Media.*

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 850–858.

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

[4] Judd Antin and Elizabeth F Churchill. 2011. Badges in social media: A social psychological perspective. In *CHI 2011 Gamification Workshop Proceedings*. ACM New York, NY, 1–4.

[5] Srinivasan Balakrishnan and Mitchell P Koza. 1993. Information asymmetry, adverse selection and joint-ventures: Theory and evidence. *Journal of economic behavior & organization* 20, 1 (1993), 99–117.

[6] Andrew Begel, Jan Bosch, and Margaret-Anne Storey. 2013. Social networking meets software development: Perspectives from github, msdn, stack exchange, and topcoder. *IEEE Software* 1 (2013), 52–66.

[7] Kelly Blincoe, Jyoti Sheoran, Sean Goggins, Eva Petakovic, and Daniela Damian. 2016. Understanding the popular users: Following, affiliation influence and leadership on GitHub. *Information and Software Technology* 70 (2016), 30–39.

[8] Amiangshu Bosu, Christopher S Corley, Dustin Heaton, Debarshi Chatterji, Jeffrey C Carver, and Nicholas A Kraft. 2013. Building reputation in stackoverflow: an empirical investigation. In *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE, 89–92.

[9] Andrea Capiluppi, Alexander Serebrenik, and Leif Singer. 2013. Assessing technical candidates on the social web. *IEEE software* 30, 1 (2013), 45–51.

[10] Yan Chen, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review* 100, 4 (2010), 1358–98.

[11] Brian L Connelly, S Trevis Certo, R Duane Ireland, and Christopher R Reutzel. 2011. Signaling theory: A review and assessment. *Journal of management* 37, 1 (2011), 39–67.

[12] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 1277–1286.

[13] Laura Dabbish, Colleen Stuart, Jason Tsay, and James Herbsleb. 2013. Leveraging transparency. *IEEE software* 30, 1 (2013), 37–43.

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

[15] Judith Donath. 2007. Signals in social supernets. *Journal of Computer-Mediated Communication* 13, 1 (2007), 231–251.

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

[17] Denae Ford, Justin Smith, Philip J Guo, and Chris Parnin. 2016. Paradise unplugged: Identifying barriers for female participation on stack overflow. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 846–857.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, NY, USA:.

[19] Jordan Frith. 2013. Turning life into a game: Foursquare, gamification, and personal mobility. *Mobile Media & Communication* 1, 2 (2013), 248–262.

[20] Tim Guilford and Marian Stamp Dawkins. 1991. Receiver psychology and the evolution of animal signals. *Animal behaviour* 42, 1 (1991), 1–14.

[21] Nicole Immorlica, Greg Stoddard, and Vasilis Syrgkanis. 2015. Social status and badge design. In *Proceedings of the 24th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 473–483.

[22] Shaili Jain, Yiling Chen, and David C Parkes. 2009. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM conference on Electronic commerce*. ACM, 129–138.

[23] Tomasz Kuśmierczyk and Manuel Gomez-Rodriguez. 2018. On the Causal Effect of Badges. (2018).

[24] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2857–2866.

[25] Anna May, Johannes Wachs, and Anikó Hannák. 2019. Gender differences in participation and reward on Stack Overflow. *Empirical Software Engineering* (2019), 1–23.

[26] Tobias Mutter and Dennis Kundisch. 2014. Behavioral mechanisms prompted by badges: The goal-gradient hypothesis. (2014).

[27] Oded Nov, Ofer Arazy, and David Anderson. 2014. Scientists@ Home: what drives the quantity and quality of online citizen science participation? *PloS one* 9, 4 (2014), e90375.

[28] Chris Parnin, Christoph Treude, Lars Grammel, and Margaret-Anne Storey. 2012. Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow. *Georgia Institute of Technology, Tech. Rep* (2012).

[29] Jeffrey A Roberts, Il-Horn Hann, and Sandra A Slaughter. 2006. Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects. *Management science* 52, 7 (2006), 984–999.

[30] Michael Rothschild and Joseph Stiglitz. 1978. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. In *Uncertainty in economics*. Elsevier, 257–280.

[31] Vandana Singh, Michael B Twidale, and David M Nichols. 2009. Users of open source software-how do they get help?. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE, 1–10.

[32] Michael Spence. 2002. Signaling in retrospect and the informational structure of markets. *American Economic Review* 92, 3 (2002), 434–459.

[33] Stack Exchange Inc. 2018. 2018. Stack Exchange Data Dump 2018-05-01. https://archive.org/details/stackexchange/.

[34] Margaret-Anne Storey, Christoph Treude, Arie van Deursen, and Li-Te Cheng. 2010. The impact of social media on software engineering practices and tools. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*. ACM, 359–364.

[35] Asher Trockman, Shurui Zhou, Christian Kästner, and Bogdan Vasilescu. 2018. Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem. In *Proc. International Conference on Software Engineering (ICSE)(ICSE). ACM.*

[36] Utkarsh Upadhyay, Isabel Valera, and Manuel Gomez-Rodriguez. 2017. Uncovering the Dynamics of Crowdlearning and the Value of Knowledge. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 61–70.

[37] Bogdan Vasilescu. 2014. Human aspects, gamification, and social media in collaborative software engineering. In *International Conference on Software Engineering, Doctoral Symposium (ICSE)*. ACM, New York, NY, USA, 646–649. https://doi.org/10.1145/2591062.2591091

[38] Bogdan Vasilescu, Alexander Serebrenik, Prem Devanbu, and Vladimir Filkov. 2014. How social Q&A sites are changing knowledge sharing in open source software communities. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 342–354.

[39] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. 2013. Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1341–1352.