Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Can Badges Foster a More Welcoming Culture on Q&A Boards?

Tiago Santos*
Graz University of Technology
tsantos@iicm.edu

Keith Burghardt, Kristina Lerman
Information Sciences Institute
University of Southern California
{keithab, lerman}@isi.edu

Denis Helic
Graz University of Technology
dhelic@tugraz.at

Abstract

Thriving online communities rely on a steady stream of newcomers to contribute new content. However, retaining newcomers has proven challenging. In this paper, we measure the success of an intervention used by Stack Exchange question-answering communities to create a more welcoming environment for newcomers. That intervention consisted in highlighting contributions by new users with a special indicator. We hypothesize that Stack Exchange's new policy would reduce negative reactions to new users and, ultimately, increase new user retention. We leverage causal modeling to assess the introduction of the so-called "new contributor indicator", and we find it did not counter user retention decline in the short- and long-terms. However, our results indicate it did reduce unwelcoming reactions towards newcomers in the short-term. Our work has practical implications for online community managers aiming to improve their onboarding processes.

Introduction

Online communities rely on a steady stream of newcomers to contribute new content in order to thrive. Retaining newcomers, however, is a challenge for many communities (Kraut and Resnick 2012). Better onboarding methods can lower barriers to entry for new users (Yazdanian et al. 2019) and improve their integration in the community (Allen 2006). Both effects serve to mitigate user churn (Yang et al. 2010; Slag, de Waard, and Bacchelli 2015) and allow communities to grow.

However, which onboarding methods are most suitable for a given community and which methods are the most effective in retaining new members? Previous work has shown that user badges can effectively steer individual behavior and incentivize participation in the Stack Exchange online question-answering (Q&A) communities (Anderson et al. 2013; Kusmierczyk and Gomez-Rodriguez 2018; Yanovsky et al. 2019). However, the effect of badges on new users, and their community-wide effect has not been fully investigated.

This work extends the line of inquiry from previous research by studying the impact of a "new contributor" indicator. On August 22, 2018, as part of an initiative to foster a more welcoming community culture (Hanlon 2019), Stack Exchange introduced this badge-like indicator, which appears in all questions and answers a user posts in the first week (and only in the first week) after her first question or answer (cf. Fig. 1). We hypothesize the introduction of this indicator would lead to:

H_n*:* higher retention of new users, and

H_c*:* fewer unwelcoming reactions by the Stack Exchange community to their contributions.

Note our hypotheses assess the new contributor indicator in terms of its impact on new users (do they churn less?) and on the community (does it become more welcoming?). We focus on *unwelcoming* reactions, as Stack Exchange identifies those as one of the main contributors to an unfriendly community culture (Hanlon 2019).

We measure the effects of the indicator in the short-term (within one month of the introduction) and long-term (within five months of the introduction). We control for long-term temporal trends in the community behavior by using a difference-in-difference regression.

In our experiments, we do not find evidence for H_n in the short-tern, nor in the long-term, but we do confirm H_c in the short-term. Our work thus sheds light on the causal effects of a new user indicator in counter-acting strong community trends. We conclude by reflecting on how community managers can capitalize on the short-term impact of badges for newcomers and thereby improve user onboarding.

Methodology

Data. We study the Stack Exchange network of community question-answering websites[1], a total of 168 communities with millions of users asking and answering questions on topics ranging from astronomy to writing. Stack Overflow, the largest (with eleven million users) and oldest community (online since 2008), is dedicated to questions related to programming. Although we focus on Stack Overflow, we also analyze all other Stack Exchange communities. We obtained a snapshot of the complete Stack Exchange network

*Research done during an internship at the Information Sciences Institute, University of Southern California.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] https://stackexchange.com/sites

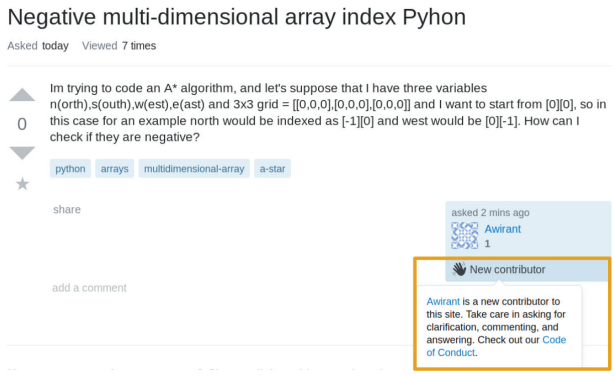

Figure 1: New contributor indicator example. In the first week after a user's first question or answer, the new contributor indicator highlighted in orange below her username encourages other users to mind their interactions with her.

as of May 2019[2].

Hypothesis Measurement. In this work, we study the effect of the introduction of a measure aimed at improving the onboarding process in Q&A communities, namely an indicator marking contributions by new users (see Fig. 1). We assess the effect of this new indicator by testing the hypotheses (H_n) new user retention increases, and (H_c) unwelcoming reactions by the community to new users decrease.

To test H_n, we define P_n as the proportion of new users who contribute a question or answer at least once within a fixed time period that extends from three days to three months after the first question. The range of the time period starts at three days to avoid including one-time users, who only return to the site to ask and follow-up on a single question. These users will not become active contributors to the community. Similarly, we set the range of the time period to end at three months to exclude users who are rarely active and are not significant contributors to the community. Our findings are robust to variations of this time window.

To test H_c, we define P_c as the proportion of first questions asked by new users which receive at least one comment with negative sentiment, based on the sentiment analysis tool VADER (Hutto and Gilbert 2014). We concentrate on the language of comments, as they provide a platform known for unwelcoming reactions (Silge and Punyon 2019). Following VADER's recommendation[3], we consider sentiment ≤ -0.05 to be negative.

Preprocessing. For both hypotheses, we focus on the first question (rather than first answer) asked by newcomers, since close to 80% of new users start by first asking questions rather than answering others' questions.

In our analysis of H_c, we consider only comments written within the one-week period after a user's first question, as that is time-frame of the "new contributor" indicator. We also separate first questions by the number of comments they obtained, test each of them separately, and report individual and aggregate results. This is because the probability of a negative comment naturally rises with more comments. We focus on first questions which receive between one and ten comments (97.5% of all first questions).

Experimental Setup. We assess whether our hypotheses hold in the *short-term*—a two-month window centered on August 22, 2018—and in the *long-term*—a ten-month window centered on August 22, 2018. Changing the short-term window to one (resp. four) months slightly reduces (resp. increases) the magnitude but not the statistical significance of our results and interpretations. The long-term window corresponds to the longest time-frame our data affords. In total, our short-term analysis of H_n comprises $60\,222$ new users, and the long-term one $562\,357$, whereas the short- and long-term analysis of H_c contains $36\,047$ and $364\,128$ first questions, respectively.

For the short-term window effects, we report the 95% bootstrapped confidence intervals (CI) for the weighted percent change in the levels of P_n and P_c. However, seasonal and other temporal trends could potentially distort the measurements of the long-term effects (Oktay, Taylor, and Jensen 2010). To control for such trends, we perform a difference-in-difference analysis (Abadie 2005). Specifically, we compare changes around August 22, 2018 to the same period in 2017, since there was no new contributor indicator then. We fit a linear model to the weekly time series of the metric values P_n and P_c:

$$P_i \sim I_{2017} + I_{Intervention} + Week \quad (1)$$

where I_{2017} is an indicator for the year 2017, $I_{Intervention}$ is an indicator for before or after August 22, 2018, and $Week$ stores the week of the year. Borrowing terminology from the causal inference literature, we name the year 2017 the *control* and 2018 the *treatment*. We inspect the magnitude and significance of the model's coefficients to assess the indicator's effects.

We also experimented fitting a logistic regression with the same regressors but a variation of P_i without temporal aggregation. In this alternative model, we let P_i be 1 for each user (resp. first question) which churns (resp. receives an unwelcoming reaction) as previously defined, and 0 otherwise. Although this model has a higher granularity, we obtain very similar quantitative results at a comparable statistical significance. For visualization purposes, we report on the weekly time series model only.

Finally, we extend both short- and long-term analyses of Stack Overflow to the other 168 communities of the Stack Exchange network. In this process, we exclude Stack Exchange communities too young to have the two years of data our long-term analysis requires, as well as those with fewer than an average of 100 new users and new first questions to analyze. That threshold corresponds to excluding communities with less than 0.2% of the newcomer activity we observe in Stack Overflow in the short-term. This leaves us with 50 communities ($\approx 30\%$ of all communities) to analyze H_n and 34 communities ($\approx 20\%$ of all communities) for H_c. We

[2]Data source: https://archive.org/download/stackexchange/.

[3]https://github.com/cjhutto/vaderSentiment#about-the-scoring

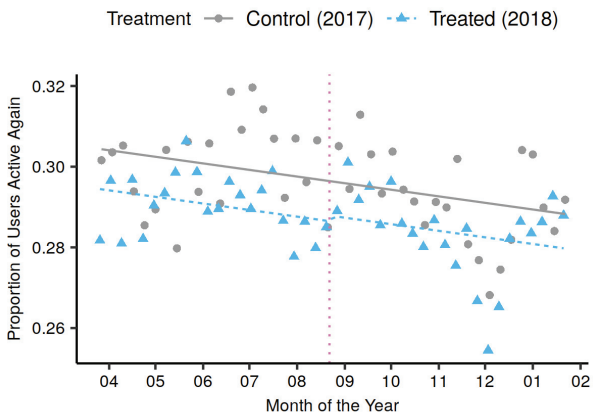

Figure 2: No significant long-term change in user retention. Controlling for temporal trends via a difference-in-difference regression on a user retention measure, we observe an overall downwards trend, which the introduction of the indicator (marked with a pink dashed line) did not curb.

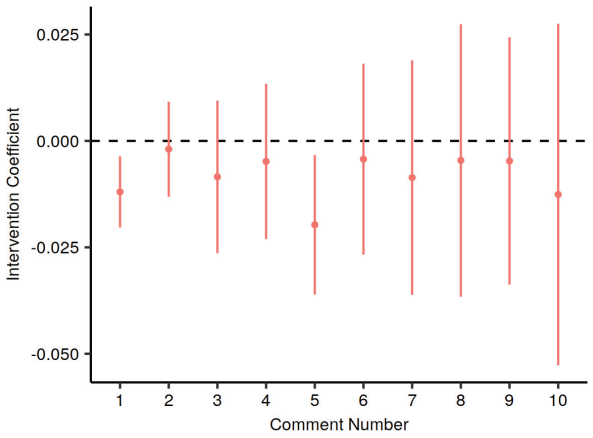

Figure 3: Few significant long-term changes in unwelcoming reactions. We control for temporal trends via a difference-in-difference regression on a measure for unwelcoming reactions. In almost all regressions studied, the intervention coefficient $I_{Intervention}$ is not significantly different from zero.

conservatively correct for repeated testing of our hypotheses by considering statistical significance at the Bonferroni-corrected p-value of $0.05/168 = 0.000298$.

Results

We first describe short- and long-term measurements of H_n and H_c on Stack Overflow, and then both time horizons of both hypotheses on other Stack Exchange communities.

Short-Term Effects. We observe small but statistically significant changes. Both user retention (P_n) and unwelcoming reactions (P_c) appear to decrease slightly. Specifically:

> H_n: P_n changed by -0.91% (CI: $[-1.28\%, -0.55\%]$, bootstrapped $p < 0.0001$).
>
> H_c: P_c changed by -1.13% (CI: $[-1.63\%, -0.64\%]$, bootstrapped $p < 0.0001$). Testing this hypothesis separately for first questions with different numbers of comments also resulted in significant changes with mostly the same magnitude and direction: In questions with between one and four comments (78% of all first questions), the weighted change is -1.65% and significantly different from zero (all four bootstrapped $p < 0.002 <$ Bonferroni-corrected $p = 0.05/10 = 0.005$). In questions with between five and ten comments (remaining 22%), the weighted change is 0.0075% and non-significant (all six bootstrapped $p > 0.13$).

When repeating this analysis for data in the year 2017, we find a statistically significant difference of similar magnitude in P_n, but no significant difference in P_c.

Long-Term Effects. We do not observe statistically significant changes as measured by the magnitude and significance of the $I_{Intervention}$ regressor. In the regression for H_n, the statistically significant downward trend of the regression persists in both years, potentially confounding the short-term estimates. On the contrary, the regression for H_c does not feature such a long-term trend. Specifically:

> H_n: The coefficient of $I_{Intervention}$ is 0.13% and it is not statistically significant (t-test $p = 0.72$), in contrast to all other regressors (t-test $p < 0.0006$). We depict this regression in Figure 2.
>
> H_c: Fitting separate regressions for first questions with different numbers of comments yielded mostly no significant coefficients (cf. Fig. 3; almost all t-test $p > 0.35$). The sole exception is the intervention coefficient of $I_{Intervention}$ in the regression for questions with one comment (t-test $p = 0.00499 <$ the Bonferroni-corrected p-value $0.005 = 0.05/10$). In particular, none of the separate regressions has a significant coefficient $Week$ for the temporal trend (all t-test $p > 0.11$ or remarkably larger), as exemplified in Figure 4.

Effects in Other Communities. In the few communities with enough data to analyze H_n, we find evidence supporting the hypothesis (after Bonferroni correction) of the short-term effect in only one of the communities, Software Engineering. The magnitude of the change is 5.36% (CI: $[1.79\%, 8.93\%]$, bootstrapped $p < 0.0001$), and there is no significant short-term change in 2017. However, in Software Engineering, both the effect and trends are absent in the long-term (in contrast to Stack Overflow), as none of the regression coefficients of our difference-in-difference analysis are significant. Regarding H_c, we find five communities (English, Mechanics, Travel, Android, and Worldbuilding) which benefit from the introduction of the new user indicator in the short-term (again, changes are statistically significant after Bonferroni correction and not present in 2017). The

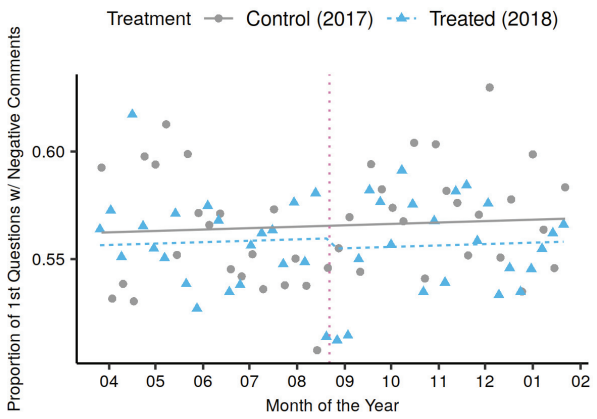

Figure 4: Example of temporally stable community reactions to newcomers. This difference-in-difference regression with only first questions which attracted four comments exemplifies the non-significant trend ($Week$) and intervention ($I_{Intervention}$) coefficients we observe across questions with varying numbers of comments.

change magnitudes range from -10.86% to -5.1% (bootstrapped $p < 0.0001$). Here, we again find no significant $Week$ or $I_{Intervention}$ coefficients for all H_c long-term regressions, similarly to our H_c results with respect to Stack Overflow.

Discussion

We inspected the effect of the introduction of an indicator to mark contributions by new users of Q&A communities. We do not find evidence for H_n (the indicator increased retention of new users) in the short- and long-terms, but our results support H_c (the indicator reduced unwelcoming reactions to new users) in the short-term.

On Stack Overflow, there is an ephemeral effect in the community's response to the new contributor indicator, but it does little to stem the long-term decline in new user retention. In particular, notice that P_n declines slightly in the short-term, but, in the long-term (see Fig. 2), there is a small but statistically non-significant positive change associated with the intervention. We reason this discrepancy arises from contextualizing fluctuations of the short-term estimate in the long-term: The decrease in August 2018 is not as strong as the overall long-term decline, and hence we observe a non-significant upwards jump slightly countering the significant downwards trend in new user retention. This result contrasts with our measurements for H_c, where none of the long-term trends are statistically significant. This, in turn, substantiates our finding that there is a reduction in unwelcoming reactions to first questions in the short-term, and that this change subsides in the long-term. This may indicate veteran users become habituated to the new indicator. Therefore, although our results indicate positive short-term changes, enacting long-term changes may require further involvement from all community members. For example, in addition to highlighting the newcomers, community managers may want to also reward veteran users for being particularly welcoming, as well as for mentoring new users (Ford et al. 2018).

Although the smaller volume of data in other communities limits our ability to generalize our findings, our results show heterogeneous effects arising from the introduction of the indicator: While almost all communities do not register an increase in user retention resulting from the indicator, a non-negligible number of them feature significant short-term decreases in unwelcoming reactions. We also note that the topics of the communities where we measured such significant changes span a broad spectrum, indicating reactions to the new indicator may be independent of the topic. This is a surprising finding, given recent work (Dev et al. 2018; Santos et al. 2019) identified topic as a key component of other community development parameters. Thus, although this calls for further research on reactions to new badges, this finding may also encourage practitioners to deploy site-wide (as opposed to community-specific) welcoming initiatives.

Although we believe the metrics we propose to measure user retention and welcoming attitudes capture our hypotheses well, future work may leverage many other metrics. In particular, using VADER to measure comment sentiment reflects only one facet of how comments may be perceived as unwelcoming; future research efforts could aim to characterize other aspects of unwelcoming reactions to newcomers (Silge and Punyon 2019).

As in all causal inference, we cannot rule out that exogenous variables and unmeasured confounders could affect our results. For example, in H_c, the number of comments a question attracts may relate to its quality and need for commenting: Do poor first questions receive more comments? And how does answer quality change and potentially affect this intervention? It would be interesting to study how such factors and other confounding community characteristics, such as age, user mix or strength of norms (Chandrasekharan et al. 2018), may impact our findings.

Finally, extending our approach to measure the impact of badges in welcoming initiatives of websites beyond Stack Exchange would help characterize the potential and limits of badges in steering community culture.

Acknowledgments. We thank the anonymous reviewers for their valuable feedback on the manuscript. Tiago Santos is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology. The work was also supported, in part, by DARPA under contracts HR00111990114 and W911NF-18-C-0011.

References

Abadie, A. 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*.

Allen, D. 2006. Do organizational socialization tactics influence newcomer embeddedness and turnover? *Journal of Management*.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2013. Steering user behavior with badges. In *WWW*.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*.

Dev, H.; Geigle, C.; Hu, Q.; Zheng, J.; and Sundaram, H. 2018. The size conundrum: Why online knowledge markets can fail at scale. In *WWW*.

Ford, D.; Lustig, K.; Banks, J.; and Parnin, C. 2018. We don't do that here: How collaborative editing with mentors improves engagement in social q&a communities. In *CHI*.

Hanlon, J. 2019. Stack overflow isn't very welcoming: It's time for that to change. https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/.

Hutto, C., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Kraut, R., and Resnick, P. 2012. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press.

Kusmierczyk, T., and Gomez-Rodriguez, M. 2018. On the causal effect of badges. In *WWW*.

Oktay, H.; Taylor, B.; and Jensen, D. 2010. Causal discovery in social media using quasi-experimental designs. In *SOMA*.

Santos, T.; Walk, S.; Kern, R.; Strohmaier, M.; and Helic, D. 2019. Self- and cross-excitation in stack exchange question & answers communities. In *WWW*.

Silge, J., and Punyon, J. 2019. Welcome wagon: Classifying comments on stack overflow. https://stackoverflow.blog/2018/07/10/welcome-wagon-classifying-comments-on-stack-overflow/.

Slag, R.; de Waard, M.; and Bacchelli, A. 2015. One-day flies on stack overflow—why the vast majority of stackoverflow users only posts once. In *MSR*.

Yang, J.; Wei, X.; Ackerman, M.; and Adamic, L. 2010. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. In *ICWSM*.

Yanovsky, S.; Hoernle, N.; Lev, O.; and Gal, K. 2019. One size does not fit all: Badge behavior in q&a sites. In *UMAP*.

Yazdanian, R.; Zia, L.; Morgan, J.; Mansurov, B.; and West, R. 2019. Eliciting new wikipedia users' interests via automatically mined questionnaires: For a warm welcome, not a cold start. In *ICWSM*.