



Gamifying with badges: A big data natural experiment on Stack Exchange

by Benny Bornfeld and Sheizaf Rafaeli

Abstract

Badges are a common gamification mechanism used by many crowd-sourced online systems. This study provides evidence to their effectiveness and measures their effect size using a big data natural experiment in three large Stack Exchange online Q&A sites. We analyze the introduction of 22 different badge-launch events and the resulting changes in user behavior. Consistent with earlier studies, we report that most badge introductions have the desired effect. Going beyond traditional findings on the individual level, this study measures overall badge effect size on the service.

Contents

[Introduction](#)

[Badges](#)

[Stack Overflow and Stack Exchange background](#)

[Gamification framework](#)

[Related theories](#)

[Related work](#)

[Method](#)

[Results](#)

[Discussion](#)

[Conclusion](#)

Introduction

Online crowd-sourced services use different mechanisms to motivate users to contribute to their platform and to steer their behavior. Points, leaderboards and badges are the most common mechanisms. In recent years there is a growing body of knowledge, exploring the motivations and effectiveness of different mechanisms.

Several studies, outlined in later in the section [Related work](#), show that if used wisely and in the right context, badges induce motivation and can steer user behavior.

This paper examines the actual effect of badges, in a field-based, natural experiment, large-scale data project, conducted in massive online participatory arenas.

In this study we examine change in user behavior in reaction to the introduction of new badges in three large Q&A communities within Stack Exchange. In those three communities, Stack Exchange

has introduced over the years 22 new badges in an endeavour to steer user behavior towards desired goals, such as answering old posts.

The introduction of new badges formed a natural experiment, in which the size of a given badge's effect on desired behavior can be measured. This is done by comparing the desired behavior levels in a rather short time window before and after a specific badge's introduction. This method is novel in this context and offers an opportunity to measure the overall impact of different badges on desired behavior within a service.

The rest of the paper is organized as follows: We start by providing some background on badges and on the Stack Exchange service. We then present the gamification framework and related theories that aim to explain why and how badges motivate people. Next we present related work and describe the method and data source. Next we present the results, showing the different badge effect sizes, and conclude with a discussion on the advantages, limitations and generality of our findings.



Badges

Preface

Some notable examples of platforms using badges are Wikipedia, Stack Exchange, Google News, Huffington Post, Khan Academy, Codecademy, Duolingo, eBay, Amazon, TripAdvisor and Foursquare. Badges are multifaceted and hence are used in different domains and for different purposes.

Antin and Churchill (2011) present five primary functions of badges: goal setting, instruction, reputation, status/affirmation and group identification. Some of them are personal such as goal setting, and some are relevant only in social context, such as reputation and status. Different online platforms use badges for different purposes. For example, Khan Academy uses badges for goal setting, instruction and affirmation, while Stack Exchange use it also for reputation and status.

In order to understand the badge mechanism, it is essential to examine its deep off-line historical roots.

History

Halavais (2012) demonstrates the great motivation promise embedded in badges, by quoting Napoleon's historian:

“... Incited by the promise of a bit of ribbon, to be stuck in the button-holes of the bravest by their emperor, whole armies of Napoleon's soldiers rushed on to meet death.” (Blanc, 1844)

Halavias (2012) reviews the different roles of off-line badges such as identification as a member of a group (university), authority (police), privileges (Delta Airline's gold medallion), achievements (medals), levels (martial arts), skills and path (Scouts) and sacrifice (medieval pilgrims).



Stack Overflow and Stack Exchange background

Stack Overflow (SO) [1] is the world's leading Q&A Web site for programming topics. The service was founded in 2008 and is free of charge. According to Alexa [2], as of November 2016, SO's popularity ranked 44 in the world, 55 in the U.S. and 18 in India. The site contained about 13 million questions and 20 million answers. About 74 percent of the questions posted on SO were answered. The service has about 10 million visits per day and there are about five million registered users.

In 2009, following the success of SO for programmers, SO founders have founded Stack Exchange (SE) [3]. SE is a network of question and answer Web sites on diverse topics, based on the SO platform. The SE network covers topics from a wide range of fields: Technical, scientific, hobbies, religion, languages and others. In 2016, the SE network contained over 150 different sites. The different sites vary in size and traffic [4]. All other SE sites are relatively smaller than SO.

Stack Exchange makes heavy use of game elements including points, leaderboards and badges. In 2016, SE badge system offered eighty four different types of badges. Jeff Atwood, one of the founders of Stack Overflow, stated that the Xbox achievement system was the inspiration and model for the Stack Overflow achievements system [5].

SO and other communities in SE are popular test beds for research. Reasons for their appeal for researchers are the services' phenomenal success, availability of raw and metadata, ability to conduct comparative study between different SE communities, and myriad game elements and interactions.



Gamification framework

We propose to view the use and impact of the badge mechanisms in the context of a gamification framework. One definition of gamification is: "The use of game design elements in non-game contexts" (Deterding, *et al.*, 2011). This refers to elements such as badges, achievements, avatars, collections, combats, content unlocking, gifting, leaderboards, levels, points and quests.

Hamari (2013) defined the purpose of gamification as "A process of enhancing services with motivational affordances in order to invoke gameful experiences and further behavioral outcomes". Huizinga's (1950) "Homo ludens" presented the importance of play and games as elements of culture and society.

Hamari (2013) referred to gamification as the process of "Transforming homo economicus into homo ludens".

In the field of human computer interfaces (HCI), the gamification concept is rooted in Malone's (1982) heuristics for designing enjoyable user interfaces, based on lessons from computer games.



Related theories

The different related theories describe the origin and nature of motivations induced by gamification. This review presents three theories: Technology Acceptance Model (TAM) proposed by Davis, *et al.* (1992), Organismic Integration Theory (OIT) (Ryan and Deci, 2000) and Social Comparison Theory (Festinger, 1954).

According to the Technology Acceptance Model (TAM) (Davis, *et al.*, 1989), the basic two parameters which govern technology acceptance are usefulness and ease of use. Three years after postulating the TAM model, Davis, *et al.* (1992) acknowledged a third element called “Perceived Enjoyment” defined as: “The extent to which the activity of using the computer is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated”.

Van der Heijden (2004) concluded that in most settings the effect of perceived enjoyment is weaker than the effects of the two original beliefs. However, he found some exceptions, such as reports of using the World Wide Web, home and leisure environment, gaming and game-based training versions of work-related information systems.

The literature distinguishes between intrinsic and extrinsic motivations (Deci, *et al.*, 1999). For example: the utility derived from doing things for fun is considered internal, while doing things to impress others generates external utility. Intrinsic utility is usually considered more desirable and argued to have a longer lasting effect (Ryan and Deci, 2000). Rewards and punishments are mechanisms to exert external utility. Badges and other game elements can be viewed as types of rewards and will initially induce extrinsic motivation.

Ryan and Deci (2000) developed the Organismic Integration Theory (OIT). OIT depicts the different ways in which extrinsically motivated behavior is regulated. Hamari (2013) offers to see gamification as “An attempt to convert utilitarian services [to be] more hedonically oriented”. This view relates to the OIT: Since hedonic motivation is more intrinsic, shifting utilitarian motivations such as game elements rewards towards intrinsic motivations will result in higher motivation and longevity.

The Social Comparison Theory (Festinger, 1954), states that people have a need to evaluate their abilities and opinions. When evaluation through objective, non-social means is not possible, they evaluate themselves by comparison to others. This comparison drives people to act. In the gamification context, since game elements such as points and badges are typically visible to others, users compare with others and are motivated to achieve more.



Related work

Hamari, *et al.* (2014) analyzed 24 empirical studies in the gamification field. They found that the most salient tested gamification mechanisms were points, leaderboards and badges. Most quantitative studies reported partial positive effects. They conclude that in general gamification does work, but caveats exist. Two main aspects arise from many of the studies: The role of the context being gamified and the qualities of the users.

Deterding (2012) presented the view of three gamification experts. While all three practice and believe in the value of HCI gamification, they express a need for caution in design and implementation.

Anderson, *et al.* (2014) studied the effect of badges in MOOCs and found that badges increased more salient engagement.

One example of an unsuccessful endeavor was a pilot research project conducted on Wikipedia. A very light badges mechanism called Barn Stars was implemented. In 2012, Wikipedia conducted a pilot [6] to assess the impact of using a more extensive badge mechanism. They did not observe statistically significant improvements from the group of editors who received badges vs. those who did not [7]. Following the pilot's results, Wikipedia decided not to implement the extensive badge

mechanism. This is somewhat in contradiction to a study by Restivo and van de Rijt (2011) which showed that the top one percent editors are motivated by barnstars [8].

Several studies examined the effectiveness of badges in Stack Exchange. All those studies found support for significant effects on user behaviors, at least for some of the badges. Anderson, *et al.* (2013) examined whether user behavior can be altered using badges. They found that users put more effort and go out of their usual participation patterns in order to achieve a badge. They also developed a model for the problem of optimally placing badges in order to induce particular user behavior.

Grant and Betts (2013) reported an increase in users' activity just before receiving a badge. Marder (2015) found supporting evidence for the Grant and Betts study in the case of several badges and no influence for others. Cavusoglu, *et al.* (2015) reported that answers badges and question badges induce users to engage more in the answering activity.



Method

This section outlines the data corpus, the stages of data gathering and processing. Our goal is to measure the overall effect of badges on the desired behavior. The method is based on a big data natural experiment. The phrase 'natural experiment' describes a naturally occurring contrast between a treatment and a comparison condition (Shadish, *et al.*, 2002). Oktay, *et al.* (2010) describe the usage of natural experiments for causal discovery in social media. They demonstrate it in a policy change in Stack Overflow.

In our study, the treatment is the introduction of new badges by Stack Exchange. Twenty-two new badges launches were analyzed in three Stack Exchange communities ([Table 2](#)). Based on the Regression Discontinuity Model, We measure the treatment effect by comparing the desired behavior (*e.g.*, old posts editing) before and after the badges were introduced.

Stack exchange registered user activities

In Stack Exchange, registered users can perform the following operations: Ask questions, provide answers, vote on posts, write comments, edit posts and review edits and posts. While all registered users can ask and answer questions, voting, editing and reviewing are limited to users with high reputation scores.

[Figure 1](#) shows registered users' weekly activities in Stack Overflow. There are periodic changes, such as annual drops around New Year. There are some fluctuations, specifically salient in the voting pattern.

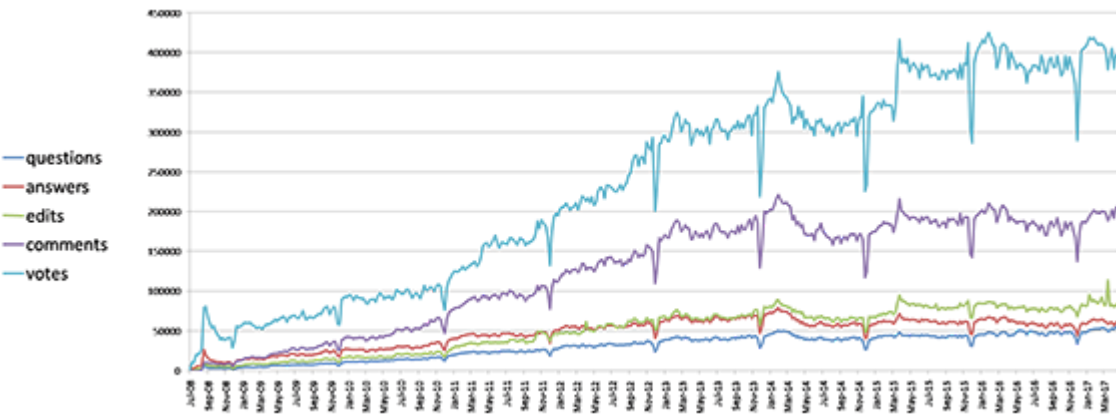


Figure 1: Weekly activities in Stack Overflow.
Note: Larger version of figure available [here](#).

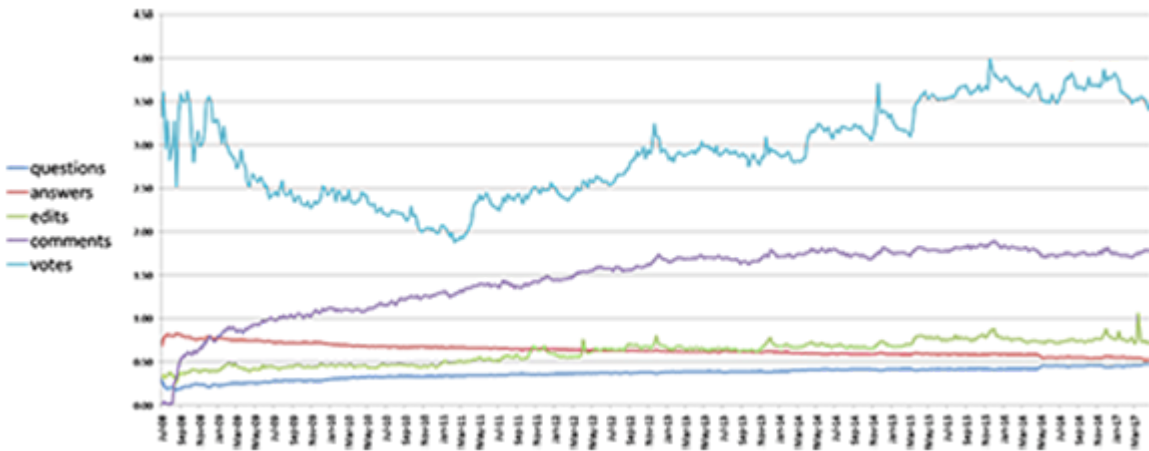


Figure 2: Weekly activities in Stack Overflow, normalized by the number of posts.
Note: Larger version of figure available [here](#).

Data corpus

Data reported here were scraped from three different Stack Exchange (SE) sites — Stack Overflow (SO), Mathematics and Super User ([Table 1](#)). Stack Overflow and Super User are arenas related to software and Mathematics contains Q&A about math. These sites were chosen for being the largest sites and among the earliest sites to be launched. Size and history are of importance in order to sustain enough data when analyzing effects on specific behaviors.

<p>Table 1: The three Stack Exchange sites used in this study (December 2016).</p> <p>Note: (*) Reached a minimum volume of 250 questions per week; (**) Active users are users who asked or answered at least one question.</p>

Site	Active since*	Number of questions	Number of answers	Number of active users**
Stack Overflow	August 2008	12,886K	20,496K	3,064K
Mathematics	November 2010	698K	989K	178K
Super User	July 2009	333K	497K	215K

Table 2: New badge description. The numbers reflect the volume of records for a period of 180 days.

Name	General description	Date	Normalized by	Stack Overflow	Super User	Mathematics
Copy Editor (Gold)	Edit 500 posts	9 July 2010	posts	96.3K	4.7K	NA
Revival (Bronze)	Answer an old question	2 November 2010	answers	26K	1.28K	NA
Excavator (Bronze) Archaeologist (Silver)	Edit 1/100 post that was inactive for six months	15 August 2011	edits	400K	9.77K	Not enough data
Explainer (Bronze) Refiner (Silver)	Edit and answer 1/50 question	30 September 2014	posts	54K	1.26K	7.57K
Custodian (Bronze) Steward (Gold)	Complete 1/1000 review task	21 September 2012	posts	691K	36.3K	19.4K
Electorate (Gold)	600 votes	2 January 2010	posts	2,340K	134K	NA
Suffrage (Bronze)	30 votes in a day	19 October 2010	posts	2,880K	117K	NA
VoxPopuli (Bronze)	40 votes in a day	9 May 2011	posts	4,410K	176K	133K
Curious (Bronze) Inquisitive (Silver) Socratic (Gold)	Asked a well received question on 5/30/100 separate days	2 July 2014	questions	382K	9.36K	46.0K

Process description

This sub-section details the techniques used in each of the study stages. [Figure 3](#) illustrates the stages of this process.

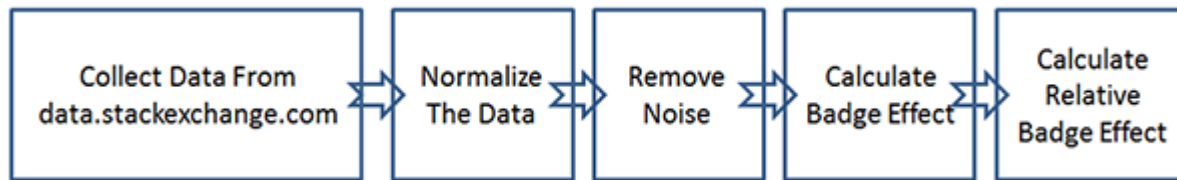


Figure 3: A diagram of the methodological process in this study.

Note: Larger version of figure available [here](#).

Data collection

The data about SE users and activities is available online at data.stackexchange.com. The data was queried using TSQL commands. Most content and data attributes are publicly available but not all. Sensitive data such as specific user's voting and reviewing data is not available.

All the data were grouped by daily amount and exported to a spreadsheet.

Data processing

Four data processing stages were employed: normalization, noise removal, calculating the badge effect and the relative badge effect ([Figure 3](#)). Analysis was conducted using R programming. In this section, we start with describing the different stages and demonstrate the process by elaborating the analysis of the Excavator and Archaeologist new badge introduction event.

Data normalizing

When studying time related phenomena, data normalization can be used to eliminate the effects of overall activity growth, periodic changes and fluctuations. Since most activities relate directly to posts, the natural choice is to normalize the different activities by posts. [Figure 2](#) shows the normalized activities for Stack Overflow. The normalized graph show that edits and comments have a linear correlation with posts and that voting is more volatile. The normalized activities for Super User and Mathematics present similar activity patterns.

In our analysis, for uniformity, simplicity and clarity, we stick whenever adequate to normalizing with posts. However, when analyzing for badges awarded for specific activities, other, more adequate normalizations are used. For example, when examining the number of answers provided for "old questions", normalization with the number of all answers is more adequate. The normalization method is specified in [Table 2](#).

Remove noise

Spikes are noises in the sense that they represent irregularities rather than normal behavior. For example, in one week in July 2012, the SO community members directed special attention and efforts at the reviews queue. During that week, the reviews count was 20 times more than the average of the previous and following months. In order to avoid impacts of spikes on the results, we

smooth them in the following manner. We define a spike as a daily value being higher or lower than twice the average of the median of its neighboring time windows (seven days on each side). We replace the spike's values with the average of the medians of the neighboring time windows.

Badge effect

The badge effect measurement is based on the Regression Discontinuity Design method (Hahn, *et al.*, 2001). The badge effect is quantified as the ratio of the average level of the desired activities before and after the badge was introduced. We used a running average window of 90 days on each side. [Figure 4](#) illustrates the way that this calculation is done. The vertical red line marks the specific date of a badge introduction. The horizontal red line at value one denotes the no change level.

The rationale for choosing a 180-day window size is the following: Making it too narrow introduces two issues. The first is higher sensitivity to local fluctuations. The second issue is that users' reaction time is not immediate since many of them are not active on daily basis. On the other hand, making the before and after window too wide introduces a pitfall. Other factors such as GUI changes, users' communications, periodic and learning effects may have affected the desired behavior.

Relative badge effect

Activity levels are constantly changing due to natural fluctuations and impacts by other factors. Within this noisy measurement there is a need to identify the signal from noise. To do so, we introduce the relative badge effect, calculated in the following manner: For each badge effect we calculate its relative size position in comparison to all the other points in the running average window ratio graph. For example, assume that the running window ratio graph for some desired activity contains values for 1,000 days. If the badge effect size calculated at the badge introduction day is higher than 950 other dates, its relative badge effect would be 95 percent. High relative effect values will increase the confidence that it is a causal change rather than some fluctuation.

To sum up, two statistics were calculated for every new badge launch event: the badge effect and the relative badge effect.

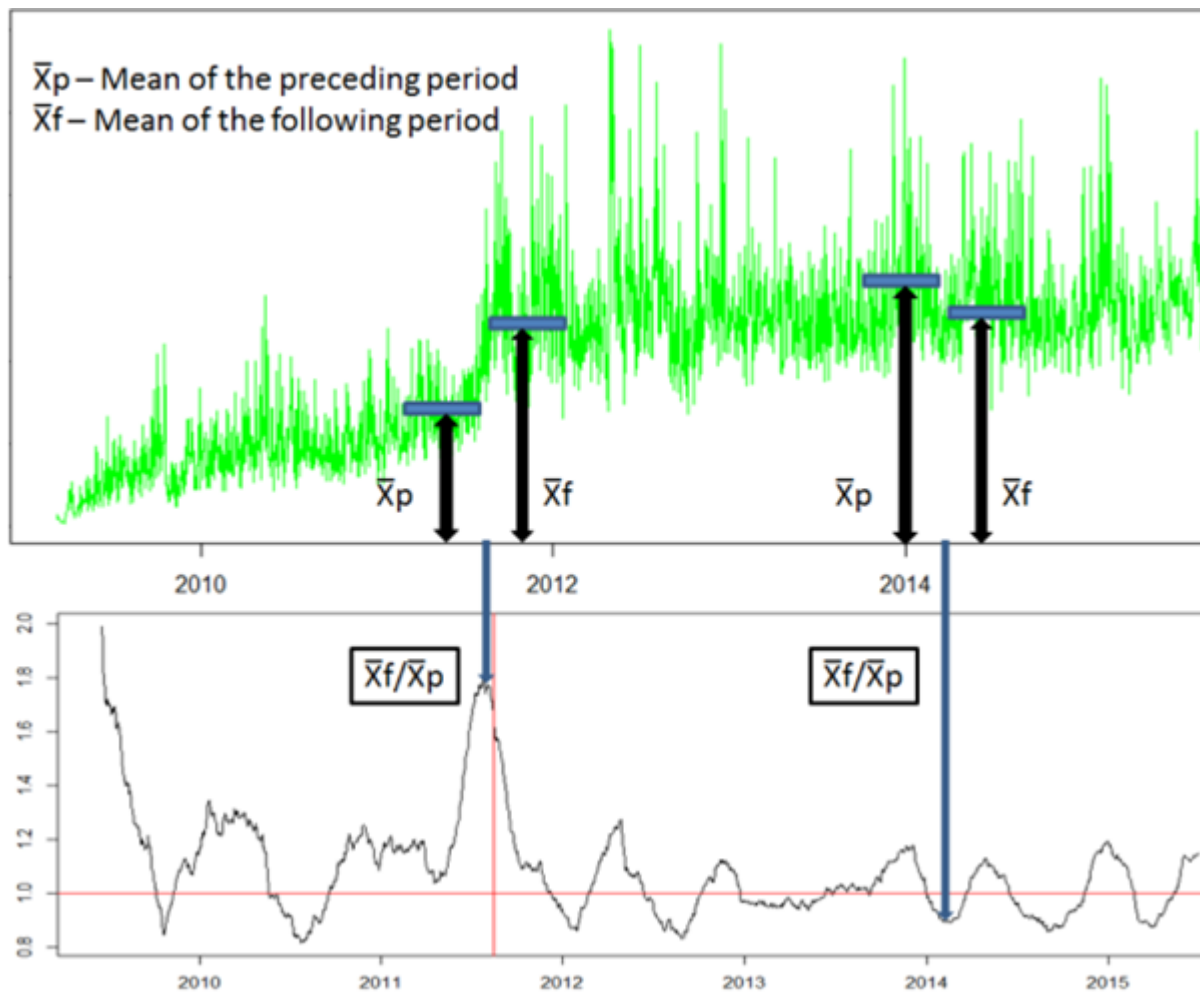


Figure 4: The running average window ratio graph.

Processing example: The Excavator and Archaeologist new badges introduction event

The Excavator and Archaeologist badges were introduced on 15 August 2011. Those badges are awarded for editing a post that was inactive for at least six months. The Excavator bronze badge is given for a single edit and the Archaeologist silver badge is won by doing a hundred edits. Both badges are awarded only once per user.

[Figure 5](#) shows the data queried from data.stackexchange.com. The yellow lines show inactive post daily editing activities along years. The blue lines show the daily number of users doing those edits. This graph suggests several observations. The amount of inactive posts edits increases over time. This can be attributed to the overall increase in editing activity. There are some high daily spikes in the amount of inactive posts editing while the number of users is more stable.

Since it measures editing activities, data was normalized by the number of edits. [Figure 6](#) shows normalized data in black. It shows that inactive posts editing grew from very low percentage at the beginning, mostly because there were no inactive posts at the very beginning of the service. Around the time of the badge introduction there was a significant increase and after that the ratio stayed more or less constant a little over 10 percent of the total number of edits.

The daily amounts were smoothed using the running average window as described earlier in the section of noise removal. The smoothed values are marked with green circles in [Figure 6](#).

[Figure 7](#) is the running average window ratio graph. The horizontal red line marks the no effect level (ratio = 1) and the vertical line marks the new badge introduction date. The analysis results are: badge effect of 64 percent, and relative badge effect of 96.6 percent.

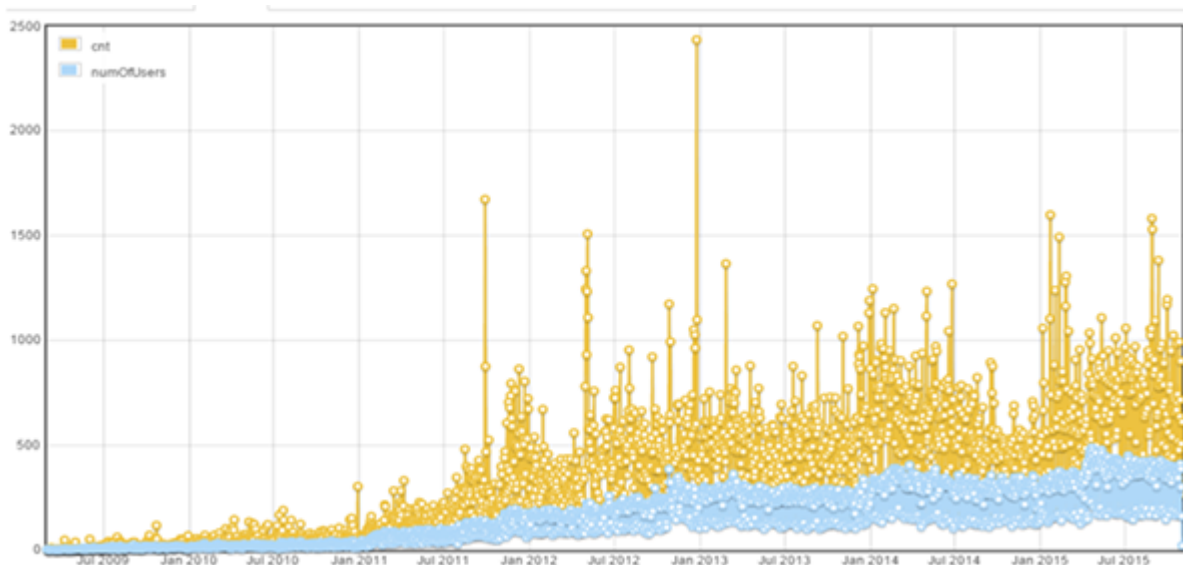


Figure 5: Stack Overflow daily data derived from data.stackexchange.com. In yellow: inactive posts editing volumes, In blue: number of users.

Note: Larger version of figure available [here](#).

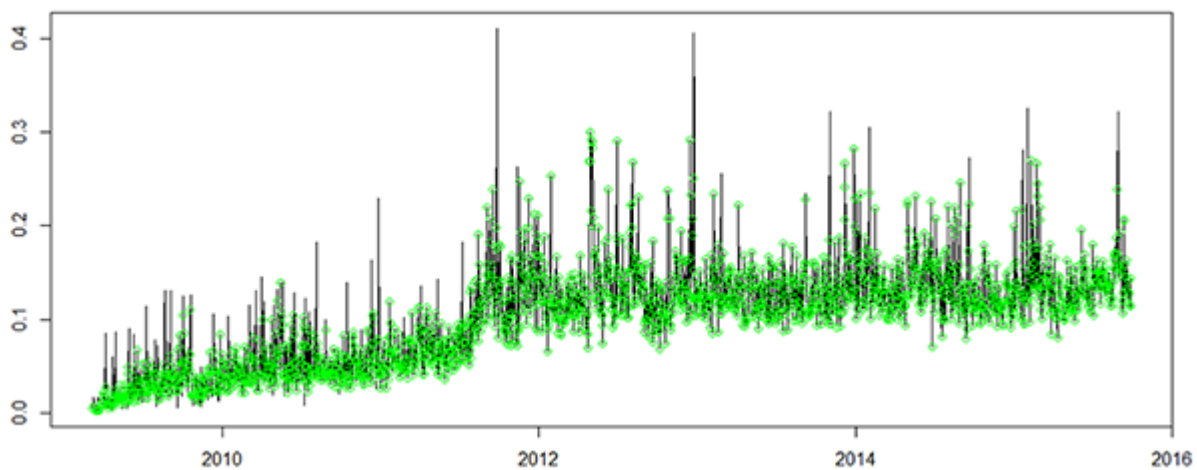


Figure 6: In black: inactive post editing normalized by edits. In green: after removing spikes.

Note: Larger version of figure available [here](#).

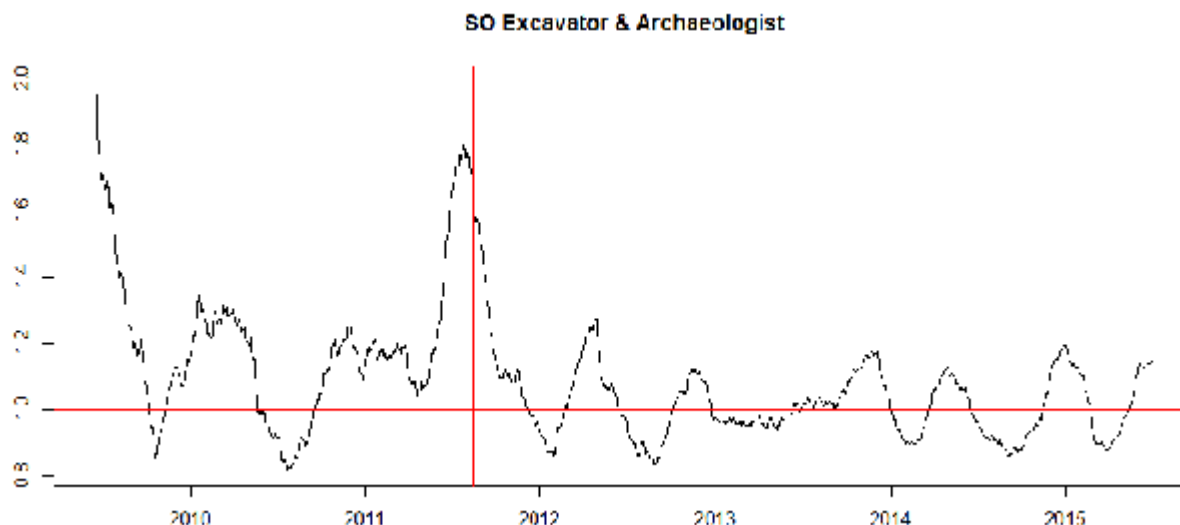


Figure 7: The running average window ratio graph, using 180-day window. The red vertical line marks badge introduction date.

Note: Larger version of figure available [here](#).



Results

We performed the above analysis for all 22 new badge introduction events in our study. The badges and event details are described in [Table 2](#). The results of the badges' effect appear in [Figure 8](#). The badge effect was positive in 18 of the 22, zero in one and negative in three events. There are considerable differences in the badges effect size. Several badge introductions, such as the Custodian and Steward, yield a large effect. Two out of the three voting related badges (*i.e.*, Electorate, Suffrage) had a minor and even a negative badge effect.

The relative badge effect results are presented in [Figure 9](#). The mean relative badge effect is 80.2 percent and the median is 88.6 percent. We tested our results statistical significance using the non-parametric Wilcoxon test. The null hypothesis was that the badges have no effect so the median would yield a mean value of 50 percent. The test yield a score $V = 238$ and $p\text{-value} < 0.01$ (0.00016), we can safely reject the no effect hypothesis.

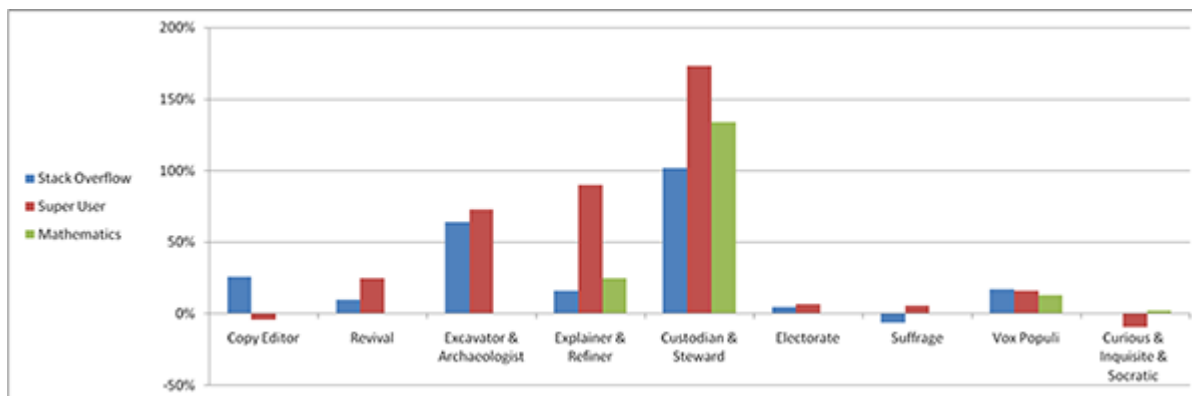


Figure 8: The badge effect: The relative change in the desired behavior following a new badge introduction.

Note: Larger version of figure available [here](#).

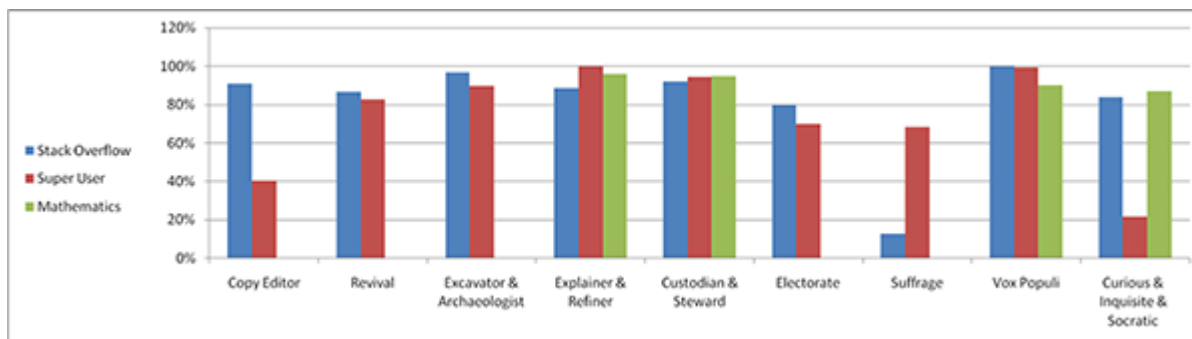


Figure 9: The relative badge effect: The badge effect relative size, compared to other fluctuations, in the running average window ratio.

Note: Larger version of figure available [here](#).

Discussion

Our findings support the hypothesis that badges have an effect on users' behavior in the desired direction. This is in agreement with previous studies (Hamari, *et al.*, 2014; Grant and Betts, 2013; Anderson, *et al.*, 2013; Cavusoglu, *et al.*, 2015; Marder, 2015).

One possible explanation for the large differences in the effect size for different badge introductions is related to the different type of users performing the desired activity. New badge introductions for reviews and posts editing yield a high badge effect in most cases. Those activities are mostly performed by high reputation, highly committed users. In activities done by all users, such as voting and questioning, the badge effect was relatively small. This is in agreement with previous studies that showed the badge effect depends on the quality of users (Hamari, *et al.*, 2014).

The method that we used presented several issues and threats to the validity of the results. We addressed them in the following fashion:


To reduce possible noise resulting from small data sets, we only used events which contained a minimum average of 50 records per week in the selected time window. In order to avoid time related issues such as holiday-period activity drop and constant growth, the data was normalized. Irregularities were handled in the noise removal stage.

Natural experiments are not conducted in a controlled environment. Other changes, such as GUI or policy changes, may have had an effect on the different activities. For example, In the case of the launch of the VoxPopuli badge, the daily limit on voting also increased from 30 to 40. We limited the analysis window to 180 days for each new badge introduction event. To control for such changes, we analyzed the feature changes log [9].

Generality limitation: the data of the different events is based on the activities of large number of users, ranging from hundreds to hundreds of thousands. However, since all the results are based on three Stack Exchange sites, those findings are limited. Many of Stack Exchange users are programmers. It can be argued that effects of badges [10] on programmers is not typical to their effects on the general population.



Conclusion

This study provides evidence for the effect of badges on user behavior. The method that we used — big data analysis of natural experimentation — for examining the overall impact of badges is novel in this context. This research implemented a method and analysis that provided quantitative results on the overall impact of selected badges in a large crowdsourced service. Those quantitative results can be used to benchmark other services considering the use of badges. 

About the authors

Benny Bornfeld is a lecturer in the Ruppin Academic Center, Emek Hefer, Israel.
E-mail: bennyb [at] ruppin [dot] ac [dot] il

Sheizaf Rafaeli is a professor at the University of Haifa and the Director of the Center for Internet Research, Israel.
E-mail: Sheizaf [at] rafaeli [dot] net

Notes

1. <http://stackoverflow.com/>.
2. <http://www.alexa.com/siteinfo/stackoverflow.com>.
3. <http://stackexchange.com/sites>; https://en.wikipedia.org/wiki/Stack_Exchange.
4. <http://stackexchange.com/sites#traffic>.
5. <http://blog.codinghorror.com/the-gamification/>; <https://www.hastac.org/documents/sociology-badges-jeff-atwood-stack-overflow>.
6. <https://en.wikipedia.org/wiki/Wikipedia:Teahouse/Badge>.
7. Personal communication, Jake Orlowitz, the project leader.
8. https://en.wikipedia.org/wiki/Wikipedia:Award_barnstars.
9. <http://meta.stackexchange.com/questions/59445/recent-feature-changes-to-stack-exchange>.
10. <https://www.quora.com/Why-many-programmers-are-avid-gamers>.

References

A. Anderson, D. Huttenlocher, J. Kleinberg and J. Leskovec, 2014. "Engaging with massive online courses," *WWW '14: Proceedings of the 23rd International Conference on World Wide Web*, pp. 687–698.

doi: <http://dx.doi.org/10.1145/2566486.2568042>, accessed 12 May 2017.

A. Anderson, D. Huttenlocher, J. Kleinberg and J. Leskovec, 2013. "Steering user behavior with badges," *WWW '13: Proceedings of the 22nd international conference on World Wide Web*, pp. 95–106.

doi: <http://dx.doi.org/10.1145/2488388.2488398>, accessed 12 May 2017.

J. Antin and E. Churchill, 2011. "Badges in social media: A social psychological perspective," *CHI 2011: Gamification Workshop Proceedings*, at <http://gamification-research.org/wp-content/uploads/2011/04/03-Antin-Churchill.pdf>, accessed 12 May 2017.

L. Blanc, 1844. *The history of ten years, 1830–1840*. London: Chapman and Hall, volume 1, pp. 555–556.

H. Cavusoglu, Z. Li and K.W. Huang, 2015. "Can gamification motivate voluntary contributions? The case of StackOverflow Q&A community," *CSCW'15: Companion Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pp. 171–174.

doi: <http://dx.doi.org/10.1145/2685553.2698999>, accessed 12 May 2017.

F. Davis, R. Bagozzi and P. Warshaw, 1992. "Extrinsic and intrinsic motivation to use computers in the workplace," *Journal of Applied Social Psychology*, volume 22, number 14, pp. 1,111–1,132.

doi: <http://dx.doi.org/10.1111/j.1559-1816.1992.tb00945.x>, accessed 12 May 2017.

F. Davis, R. Bagozzi and P. Warshaw, 1989. "User acceptance of computer technology: A comparison of two theoretical models," *Management Science*, volume 35, number 8, pp. 982–1,003.

doi: <http://dx.doi.org/10.1287/mnsc.35.8.982>, accessed 12 May 2017.

E. Deci, R. Koestner and R. Ryan, 1999. "A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation," *Psychological Bulletin*, volume 125, number 6, pp. 627–668.

S. Deterding, 2012. "Gamification: Designing for motivation," *Interactions*, volume 19, number 4, pp. 14–17.

doi: <http://dx.doi.org/10.1145/2212877.2212883>, accessed 12 May 2017.

S. Deterding, D. Dixon, R. Khaled and L. Nacke, 2011. "From game design elements to gamefulness: Defining 'gamification'," *MindTrek '11: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp. 9–15.

doi: <http://dx.doi.org/10.1145/2181037.2181040>, accessed 12 May 2017.

L. Festinger, 1954. "A theory of social comparison processes," *Human Relations*, volume 7, number 2, pp. 117–140.

doi: <http://dx.doi.org/10.1177/001872675400700202>, accessed 12 May 2017.

S. Grant and B. Betts, 2013. "Encouraging user behavior with achievements: An empirical study," *Proceedings of the 10th IEEE Working Conference on Mining Software Repositories*, pp. 65–68.

doi: <http://dx.doi.org/10.1109/MSR.2013.6624007>, accessed 12 May 2017.

J. Hahn, P. Todd and W. Van der Klaauw, 2001. "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, volume 69, number 1, pp. 201–209.

doi: <http://dx.doi.org/10.1111/1468-0262.00183>, accessed 12 May 2017.

- A. Halavais, 2012. "A genealogy of badges: Inherited meaning and monstrous moral hybrids," *Information, Communication & Society*, volume 15, number 3, pp. 354–373.
doi: <http://dx.doi.org/10.1080/1369118X.2011.641992>, accessed 12 May 2017.
- J. Hamari, 2013. "Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service," *Electronic Commerce Research and Applications*, volume 12, number 4, pp. 236–245.
doi: <https://doi.org/10.1016/j.elerap.2013.01.004>, accessed 12 May 2017.
- J. Hamari, J. Koivisto and H. Sarsa, 2014. "Does gamification work? — A literature review of empirical studies on gamification," *2014 47th Hawaii International Conference on System Sciences (HICSS)*, pp. 3,025–3,034.
doi: <https://doi.org/10.1109/HICSS.2014.377>, accessed 12 May 2017.
- J. Huizinga, 1950. *Homo ludens: A study of the play element in culture*. New York: Roy Publishers.
- T. Malone, 1982. "Heuristics for designing enjoyable user interfaces: Lessons from computer games," *CHI '82: Proceedings of the 1982 Conference on Human Factors in Computing Systems*, pp. 63–68.
doi: <https://doi.org/10.1145/800049.801756>, accessed 12 May 2017.
- A. Marder, 2015. "Stack overflow badges and user behavior: An econometric approach," *MSR '15: Proceedings of the 12th Working Conference on Mining Software Repositories*, pp. 450–453.
doi: <https://doi.org/10.1109/MSR.2015.61>, accessed 12 May 2017.
- H. Oktay, B. Taylor and D. Jensen, 2010. "Causal discovery in social media using quasi-experimental designs," *SOMA '10: Proceedings of the First Workshop on Social Media Analytics*, pp. 1–9.
doi: <https://doi.org/10.1145/1964858.1964859>, accessed 12 May 2017.
- M. Restivo and A. van de Rijt, 2011. "Experimental study of informal rewards in peer production," *PloS One*, volume 7, number 3, e34358, at <https://doi.org/10.1371/journal.pone.0034358>, accessed 12 May 2017.
- R. Ryan and E. Deci, 2000. "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *American Psychologist*, volume 55, number 1, pp. 68–78.
doi: <http://dx.doi.org/10.1037/0003-066X.55.1.68>, accessed 12 May 2017.
- W. Shadish, T. Cook and D. Campbell, 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, Calif.: Wadsworth Cengage learning.
- H. van der Heijden, 2004. "User acceptance of hedonic information systems," *MIS Quarterly*, volume 28, number 4, pp 695–704.

Editorial history

Received 1 January 2017; accepted 22 April 2017.



This paper is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

Gamifying with badges: A big data natural experiment on Stack Exchange

by Benny Bornfeld and Sheizaf Rafaeli.

First Monday, Volume 22, Number 6 - 5 June 2017

<https://journals.uic.edu/ojs/index.php/fm/article/download/7299/6301>

doi: <http://dx.doi.org/10.5210/fm.v22i16.7299>