MS&E 331 – Fall 2014
Final Paper – Tuesday December 9th
Social Dynamics on Stack Overflow

Clement Renault
Eric Boucher
Florence Koskas

The impact of Gamification on Stack Overflow user behavior

Introduction:

What's Stack Overflow

Individuals increasingly rely on their peer communities for information, advice, and expertise. Millions of people learn from each other on public discussion forums, the famous community-built Wikipedia, social networks and social Q&A (question-answering) sites. Social Q&A, according to Shah et al., consists of three components: a mechanism for users to submit questions in natural language, a venue for users to submit answers to questions, and a community built around this exchange. Stack Overflow (SO) is a social Q&A website like Quora or Answers by Yahoo but SO focuses on specialized technical knowledge: computer programming questions only. The guidelines are clear: "*Focus on questions about an actual problem you have faced. Include details about what you have tried and exactly what you are trying to do. Not all questions work well in our format. Avoid questions that are primarily opinion-based, or that are likely to generate discussion rather than answers.*"

Gamification on Stack Overflow

On SO, questions and answers are voted upon by the community. The number of votes is reflected in the users' reputation scores and badges. Exceeding various reputation thresholds grants access to additional features; also, reputation and badges can be seen as a measure of one's expertise by potential recruiters.

Problematics

With 8.5 million questions and 74% of these questions answered, SO is considered to be the biggest success in social Q&A websites. SO has 14 million answers in total which already testifies to the existence of a community of helpers. There is also a group of extremely active users on Stack Overflow who spend enormous time and effort answering numerous questions, thereby actively contributing to the community's technical knowledge. As such, a natural research questions arises: What causes people to help each other on Stack Overflow? What are the motivations of these people? In this paper, we will first give some possible answers to this question based on online users' testimonials and social science literature. Then, we will refine this question to a sub question we were able to partially explore with the data we collected from Stack Overflow: Is gamification a driver of extremely frequent participation on Stack Overflow?

I. Why do people help each other on Stack Overflow?

a. Insights from testimonials

Online testimonials suggest that Stack Overflow users are helping each other because of:

- **Reciprocity:** They think people will help them back in the future when they need it.
- **Learning objectives:** They learn by answering other users' questions as well as by getting objective criticism on why their own answers are wrong (when they are).
- **Sense of community:** They want to and enjoy being part of a community of peers
- **Building an open and enduring knowledge base:** They want to create long lasting value to a broad audience.
- **Career objectives:** A good online reputation can have a positive impact on their programming careers. Some users consider their portfolio of Stack Overflow answers a valuable component of their professional resumes.
- **Gamification:** They have a natural tendency to play and a competitive mindset and hence are motivated by the system of scores and badges.

Let's now explore some social science research on the same question.

b. Insights from social science literature

i. Reciprocity

"Pay every debt as if God wrote the bill." – Ralph Waldo Emerson
In his book "The psychology of persuasion", Cialdini defines the reciprocity rule as follows: "we should try to repay, in kind, what another person has provided us". He describes reciprocation as "one of the most potent of the weapons of influence around us". Cialdini even explains how the power of the reciprocity rule is such that by first doing us a favor, strange, disliked, or unwelcome others can enhance the chance that we will comply with one of their requests. This might suggest that people we wouldn't necessarily want to help in real life, based on personality fit only, manage to get our help in online environments like Stack Overflow by helping us in the first place. Are users active on Stack Overflow in order to get other users' help in return?

In their paper "Self-Interest, Reciprocity, and Participation in Online Reputation Systems", published in 2004, Chrysanthos Dellarocas, Ming Fan and Charles A. Wood focus on the eBay transaction platform. eBay's feedback mechanism is the primary means through which eBay promotes and maintains honest behavior and, thus, facilitates transactions between strangers over the Internet. The success of online reputation systems depends on the sustained voluntary contribution of feedback by community members. In 2004, empirical results from eBay show that buyers submit ratings to more than 50% of transactions. However, feedback submission requires effort from the providers; so, why do people provide feedback on their transaction partners on the eBay platform?
By analyzing data from 51,452 eBay rare coin auctions, the study analytically and empirically demonstrates that the expectation of reciprocal behavior from transaction partners increases the participation from eBay users in the reputation system. More generally, through theoretical and empirical analysis, the study demonstrates that the high levels of voluntary participation on eBay's reputation mechanism can be explained through the combined effects of altruism, self-interest, and reciprocation.

As such, in the eBay case, reciprocation seems to have played a role in the motivations of users for rating their transaction partners before 2004. This is supporting evidence to the "reciprocity" argument to answer our question. Let's explore now past research about gamification.

ii. Faster answers in gamified environments

In their study "How Social Q&A Sites are Changing Knowledge Sharing in Open Source Software Communities", published in February 2014, Vasilescu, Serebrenik, Devanbu and Filkov explore the changes in behavior of specific software knowledge contributors as they migrate into gamified environments.
A bit of background: R-help is a mailing list for questions and answers about problems and solutions using the programming language R, which was created in 2008. StackExchange is a website network that includes 131 different Q&A online communities modeled after Stack Overflow, which was launched in 2009. Today, StackExchange has over 89 million monthly unique visitors.
Findings: The main contribution of the mentioned study is the assembly of a joint data set from both sources, in which participants in both the R-help mailing list and StackExchange are identifiable. With this data set, they found that mailing list experts are shifting away from R-help and migrating to StackExchange, where their behavior is different. First, participants active both on R-help and on StackExchange are more active than those who focus exclusively on only one of the two. Second, they provide faster answers on StackExchange than on R-help, suggesting that they are motivated by the gamified environment.

This study shows online “helpers” provide faster answers in StackExchange, which is a gamified environment, than in R-help, which is a classic mailing list, thereby giving additional support to the gamification argument we heard in Stack Overflow user testimonials.

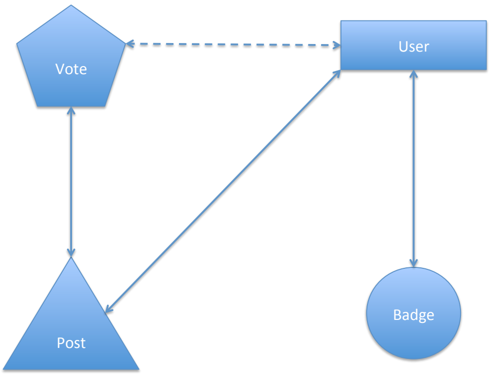
By analyzing StackExchange and eBay’s dynamics, the social science research papers described above gave supporting evidence for the reciprocity and the gamification arguments we found in Stack Overflow user testimonials. In this paper, we decided to contribute to the exploration of the question of why people help each other in online environments like StackExchange by specifically focusing on Stack Overflow. Let’s explore the data we had available, the challenges we encountered while treating the data and what we established was an interesting area of focus.

I. Data available, methodology and first insights

a. Data available and workflow

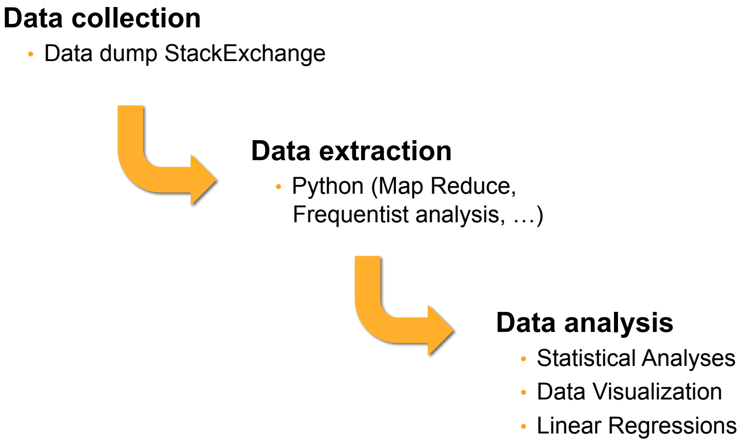
To try and answer our questions we started looking at the data available about SO users and posts. We quickly found that SO freely and regularly offers dumps of its databases. Those data dumps are available at this URL: https://archive.org/details/stackexchange.

The dumps are divided in multiple 7zipped XML files that regroup information about different categories (users, posts, badges, votes, etc.). The files that we used and their links are shown in the graph on the right. Dashed lines represent wanted yet non-existent links



Datasets used in the analysis and their links

As soon as we started working on these data dumps, it appeared that the volume of information was too large to be analyzed with traditional tools such as R, Stata and other statistical packages. Therefore we came up with a workflow that allowed us to extract meaning from these gigantic datasets. This workflow is illustrated in the following figure.

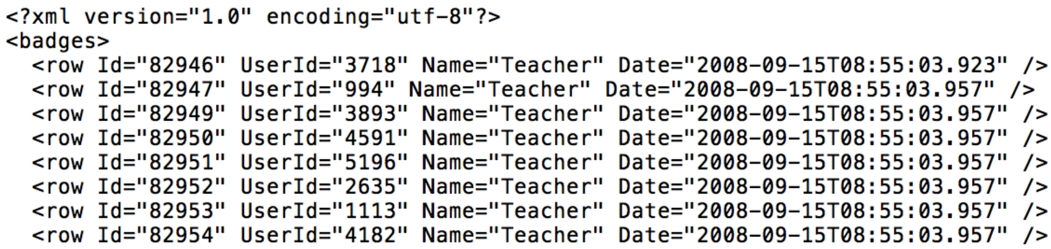


Workflow of the data analysis

Our objective was to extract as much interesting information as possible directly with Python programming language. For instance, we ran MapReduce jobs to count the average number of questions and answers per user or to extract the Id’s of users who have reached certain reputation levels. Then, when we were unclear about what exactly we wanted to analyze we created light text files with only the most interesting fields of our datasets in order to explore them more in-depth using R. The last step of our workflow was then to run analyses on these extracts and output graphs.

Even though the data consisted of relatively clean dumps directly extracted from the SO database we still faced many challenges when analyzing it. We will briefly describe one such challenge.

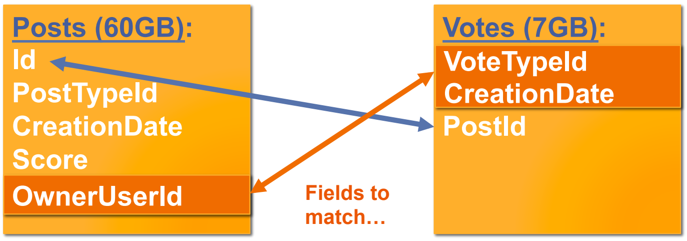
As previously mentioned, the data was divided into multiple XML files. The format of these files is shown in the figure below:



```
<?xml version="1.0" encoding="utf-8"?>
<badges>
  <row Id="82946" UserId="3718" Name="Teacher" Date="2008-09-15T08:55:03.923" />
  <row Id="82947" UserId="994" Name="Teacher" Date="2008-09-15T08:55:03.957" />
  <row Id="82949" UserId="3893" Name="Teacher" Date="2008-09-15T08:55:03.957" />
  <row Id="82950" UserId="4591" Name="Teacher" Date="2008-09-15T08:55:03.957" />
  <row Id="82951" UserId="5196" Name="Teacher" Date="2008-09-15T08:55:03.957" />
  <row Id="82952" UserId="2635" Name="Teacher" Date="2008-09-15T08:55:03.957" />
  <row Id="82953" UserId="1113" Name="Teacher" Date="2008-09-15T08:55:03.957" />
  <row Id="82954" UserId="4182" Name="Teacher" Date="2008-09-15T08:55:03.957" />
```

Datasets formatting

One of the issues we encountered was when we tried to match votes on posts to the users who made the posts. This was a hard task because the Id of the user who posted was not available in the votes databases leading to the situation shown below.



Matching votes to users: a difficult task

To solve our problem we created a MapReduce job that mapped the PostIds in the Votes dataset to the Id field in the Posts dataset. This step enabled to output the Id of the user who posted and the date the vote was recorded as the key and the vote type as the value. Then the reducer calculated the total reputation received on a given day by a specific user.

The fact that the data came directly from SO gives us a high confidence in its accuracy and therefore reinforces the strength of our findings. Moreover the use of modern computational methods allowed us to benefit from the vast datasets made available by the StackExchange community without having to sample it. This reduces the margin of error of our results.

b. General findings around Stack Overflow

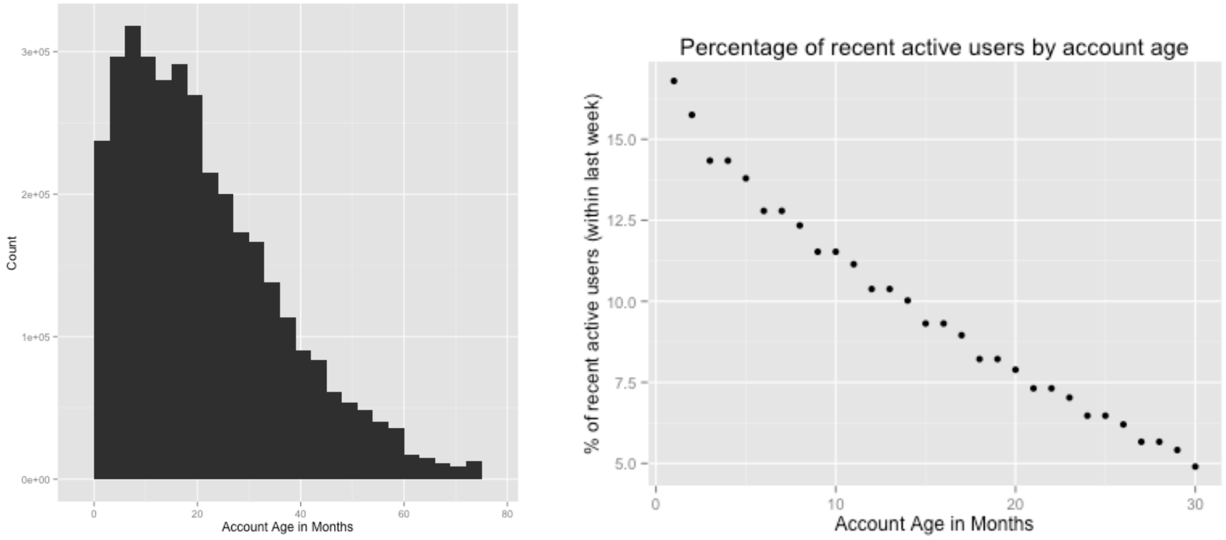
We first performed a broad analysis of Stack Overflow's user base.

There were ~3.5M users registered on SO as of September 2014. Out of them only ~50% have posted at least one post over their entire membership time (~1.9M). We still have not figured out why these users have signed-up if they did not intend to post (maybe they were interested in features such as bookmarking or maybe they thought they would become active and then never made the leap for some reason).

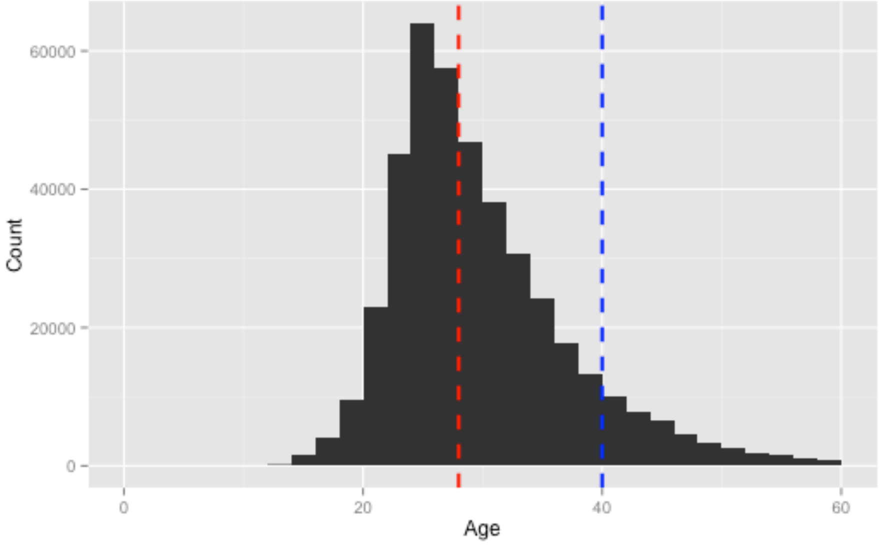
We explored user acquisition. We plotted the number of users per membership time (time since sign up) and realized that SO has gone through a phase of exponential growth after until the c. 18 months ago. It also looks like the number of users signing-up has flattened out in the past 18 months.

We also explored retention. To do so, we plotted the average % of users who logged in the past week (more precisely: the week before the data was extracted) for each membership time range – from the newest to the oldest cohorts. It clearly appears that the older the cohort the less people have logged in over the most recent week. Old users might need to be reactivated.

Finally, we looked at the age distribution of users on SO. Interestingly the median age on SO is 28 years old. This is to be compared to the median age in developed countries of 40 years old.



Number of users per account age in months (left) and percentage of users who logged in the week before the data extraction per account age in months (right)



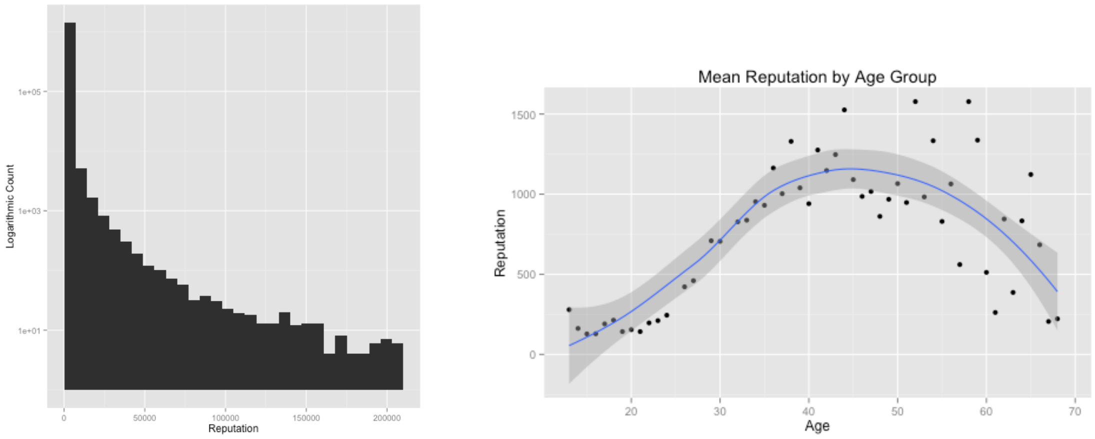
Age distribution of users on Stack Overflow. Median age on SO is represented by the red line. Median age in developed countries is represented by the blue line

c. Reputation as a proxy for "helping one another"

To simplify our study, we decided to use reputation as a proxy for being helpful on SO. SO offers a system that allows to upvote and downvote questions, answers and comments. These votes are converted into numerical scores that alter one's "reputation". The most upvotes you get the highest reputation you have. The interesting thing about the reputation score is that it is a better proxy for being helpful than posts for instance since you could post many answers that help no one. With reputation the users themselves express how valuable your posts are.

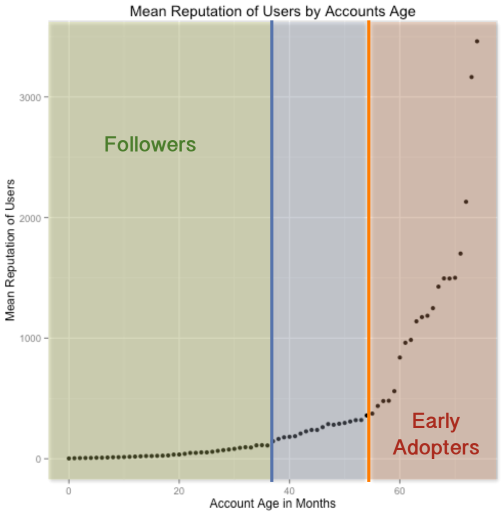
There are many rules to gain or lose reputation on SO. We will not go through them all in this paper but the interested user could find them all on this page: http://stackoverflow.com/help/whats-reputation

We looked at the distribution of reputation over registered users. It appears that 60% of registered users have no reputation. Also, less than 10% of users have a reputation above 70, which is equivalent for example to 7 answers voted up throughout the time on SO and is also 10,000 times lower than the highest reputation on SO. Another interesting fact is that, although the median age of registered users is 28 years old, the age group with the highest average reputation is c. 45 years old.



Distribution of reputation of SO users (left) and mean reputation by age group (right)

Finally, we plotted the average reputation by membership duration (time since sign up). It shows very clearly that different types of users joined at different times. The first users to sign-up are much more active and involved than recent users. This is expected as early adopters usually have a higher sense of community and tend to feel like they are almost part of the founding team.

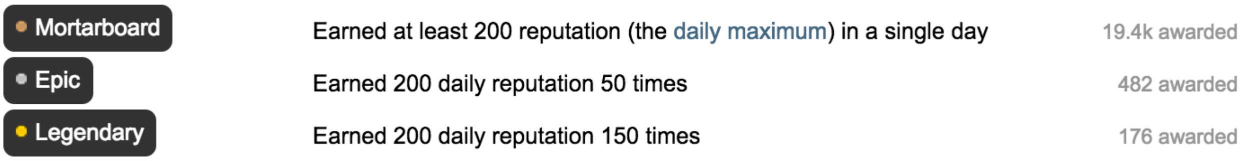


Mean reputation by time since sign up

II. The sub-question we explored: Is gamification a driver of extremely frequent participation on Stack Overflow?

a. Legendary badge as a proxy for extremely frequent participation

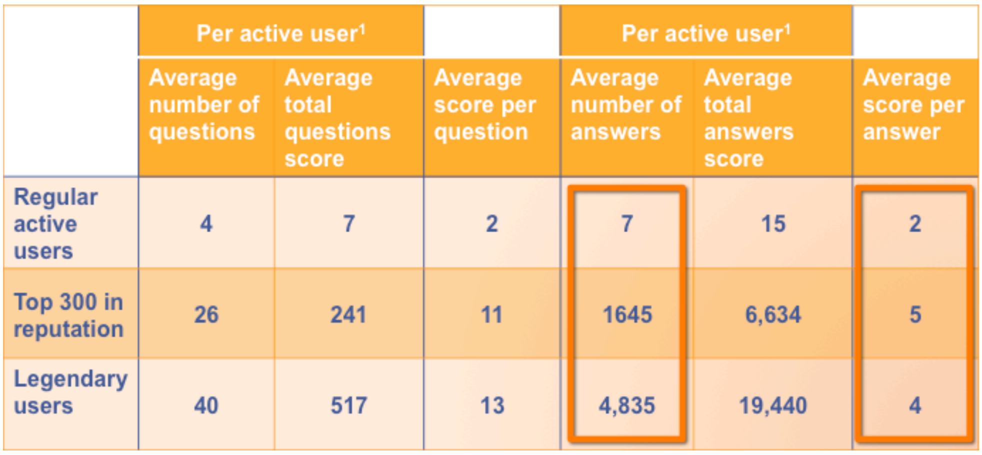
Stack Overflow has developed a badge system to reward particularly active and helpful users. Once a user gets a badge, it appears on its profile page, flair and posts. There are badges for everybody: questions badges, answers badges, comments badges, posts badges, tag badges, moderation badges, loyalty badges, reputation badges and more. Some of the most renowned reputation badges are:



● Mortarboard	Earned at least 200 reputation (the daily maximum) in a single day	19.4k awarded
● Epic	Earned 200 daily reputation 50 times	482 awarded
● Legendary	Earned 200 daily reputation 150 times	176 awarded

Based on the data that we had, we decided to use the Legendary Badge as a proxy for extremely frequent participation. Indeed, obtaining this badge requires a high level of involvement and only 174 of the all-time users of the website managed to get it. To get it, they must gain 200 daily reputation points on 150 occurrences. 200 reputation points approximately correspond to 20 of their answers being up-voted on a single day. Although they are representing only 0.005% of the user base, they contribute to more than 4% of posts. Let us compare them to other active users and try to understand what they are doing differently.

b. Comparing Legendary vs. other active users on the platform



	Per active user[1]			Per active user[1]		
	Average number of questions	Average total questions score	Average score per question	Average number of answers	Average total answers score	Average score per answer
Regular active users	4	7	2	7	15	2
Top 300 in reputation	26	241	11	1645	6,634	5
Legendary users	40	517	13	4,835	19,440	4

Comparison Table - Legendary Users vs. Other Active Users

We define an active user as someone who posted at least one question or one answer on Stack Overflow. As we could have guessed, Legendary users differentiate by a very high average number of answers posted per user with more than **4,800** compared to **7** for other active users (c. 700x more). However, it is interesting to note that the average score per answer of Legendary users is only twice as high as the average score per answer of other active users. It's not about "posting much much better quality answers", it's just about "posting much more". However, it is not certain that the average score per answer accurately reflects their quality. Indeed, only few answers to a question are up-voted and getting upgraded to the top of the feed. As a result, good answers, which were not up-voted early, might not be visible enough, not get any votes and thus, not bring any points to their author. It might be interesting to analyze this issue further.

Without any doubt, Legendary users are highly involved on Stack Overflow. But does gamification, through the badge system, influence this involvement? Does receiving a Legendary badge cause a decrease in participation for that user?

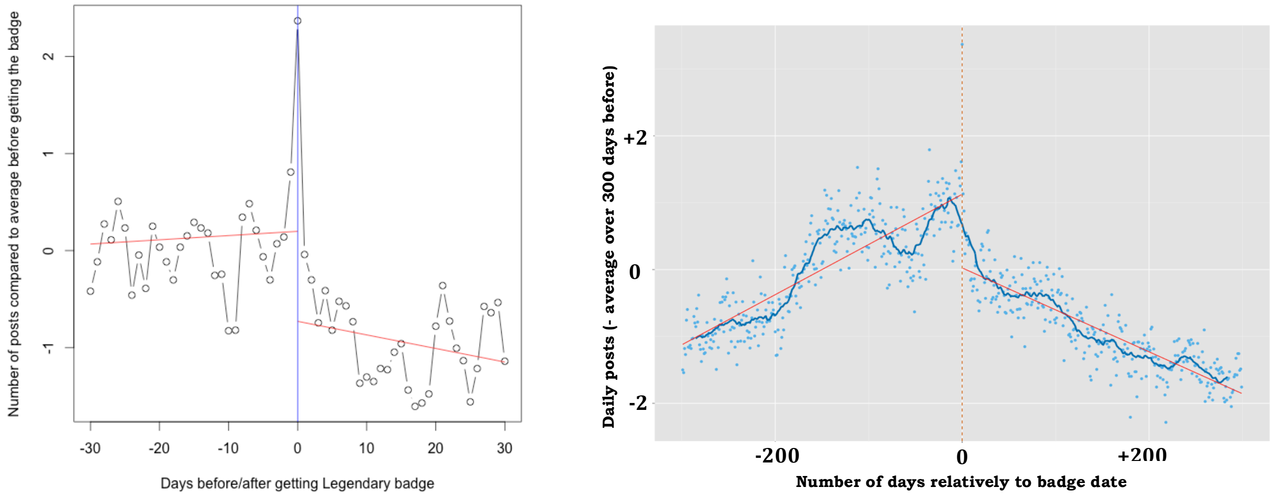
c. Does receiving a Legendary badge cause a decrease in participation?

To answer this question, we first analyzed a quasi-experimental design: the interrupted time-series design[1]. In the interrupted time series design, we observe an outcome variable for a certain time interval, Δt, before and after treatment for the same group of people (no control group). Here, we can observe the average number of posts per day before and after getting a Legendary badge for example. However, this observation over Δt lets us identify intrinsic trends within the time-series and therefore rule out some threats to validity. The causal question we consider here is:

Does receiving a Legendary badge cause a decrease in participation for that user?

We replicated the methodology followed by Hu̇seyin Oktay, Brian J. Taylor, David D. Jensen: *"The time-series will be the number of posts per day by the user with the outcome measure being the change in the number of posts by that particular user before and after treatment. Using the number of posts by the user instead of the reputation points they obtained as our outcome metric because number of posts is only influenced by the user and not by other users (voters) or external events. We therefore believe that number of posts is a better metric for user behavior. Threats to validity of the interrupted time-series include historical effects. [This risk is explained further in the IV.a. Limitations paragraph. This risk is mitigated by the fact we observe 174 different users who have potentially 174 different badge dates]. There are [174] users with the [Legendary] badge in our dataset. For each user, we determine the relative time at which they get the [Legendary] badge. Then we calculate the number of posts corresponding to each user for 30 days before they get the [Legendary] badge and 30 days after they get the badge. To make the analysis clearer, we calculate the average number of posts for the 30-day period before the treatment for each user, and we normalize the daily number of posts for each user by subtracting that average. We then calculate the average number of posts from those normalized values among [174] users for each day and plot those values. We fit linear models to the data points before and after the treatment."*

We then replicated the experiment for a time interval Δt of 300 days.

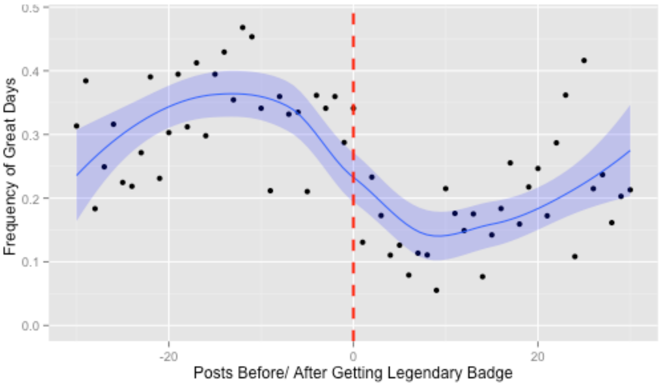


Interrupted time-series around the event: getting a Legendary badge

[1] Methodology inspired from: Causal Discovery in Social Media Using Quasi-Experimental Designs by Hüseyin Oktay, Brian J. Taylor, David D. Jensen

For Δt = 30 days, there seems to be a clear discontinuity around the day users are getting the badge. Observing a significant slope change in both cases, we can conclude that the treatment has an effect and that users tend to contribute less after getting the Legendary badge. Interestingly, this result appears to stand for Δt=300 days, suggesting long-lasting effects. At this point, it is important to remember that the Legendary badge is the highest reputation badge you can get on Stack Overflow. Does this mean that Legendary users are working hard towards the badge reward and then relax their efforts?

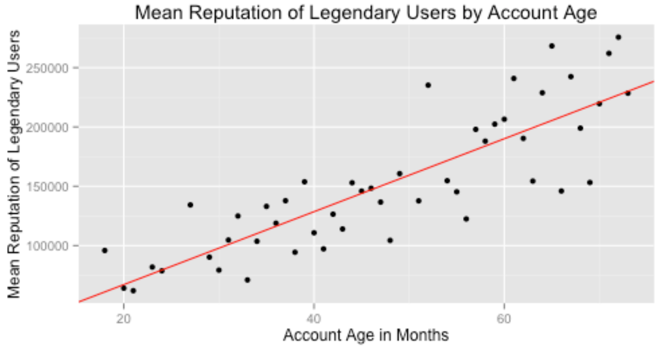
Frequency of “great days” around the event: getting the Legendary badge



We also analyzed the frequency of “great days” among Legendary users, again in an interrupted time-series fashion. We defined a “great day” as a day when the user reaches a daily reputation of more than 200 points. This is an interesting metric for Legendary users since they achieved 150 “great days” in order to get their Legendary badge.
Results: The rise and drop in daily posts (observed previously) is also noticeable in the frequency of great days around badge date (in an interval of c.10 days before and after badge date). However, the frequency seems to increase again starting 10 days after badge date.
How to explain that rise given that the daily number of posts continues to decrease at that point according to the previous analysis?

- This could suggest that owning a Legendary badge enables a Legendary user to naturally attract more up votes per post; this might be because users tend to consult more posts provided by Legendary users (and as such are more exposed to them and able to recognize their qualities), or because they subconsciously tend to have a more positive opinion of content posted by Legendary users, or both. In this scenario, this increase in frequency could actually be one of the drivers of the decrease in number of posts in the long run itself: indeed, if Legendary users are only motivated by being among the users with highest reputation, seeing their relative reputation continue to increase whereas they don’t post as much might demotivate them from posting as much as they used to.
- The rise in frequency could also simply be because the number of users on Stack Overflow increases with time and so does the number of up votes in general.

d. Can we split Legendary users between early adopters and followers?



As we observed in Part II, plotting the “reputation” as a function of “time since sign-up” allowed us to split users into three groups. Can we also find such differences among Legendary users? Is it possible to distinguish between early adopters and followers when looking at their reputation? In fact, the result is very different from the plot on all users and it does not seem to be the case. Reputation seems to linearly increase with time since sign-up, suggesting as far as Reputation is concerned, Legendary users seem to be one of a kind.

III. Limitations of the study and next steps

a. Limitations of the study

- We were missing some pieces of information like time stamped reputation data, which would have allowed us to perform a more thorough and accurate analysis.
- Also, we didn't analyze the statistical significance of our results, which potentially reduces the relevance and impact of our findings, especially given that in the interrupted time-series design the difference in daily posts "before and after" is around 1 or 2 posts.
- One of the limitations of our interrupted time-series analyses is that we don't have a comparison group which would allow us to account for other unobserved effects. As such, the decrease in the daily number of posts (or frequency of "great days") might be due to other reasons like:
    - StackOverflow-specific reasons such as: the service goes down just after the badge date. We expect these effects to average out over the potentially 174 different dates on which users got their legendary badges.
    - User-specific reasons such as: the user becomes busier in his offline life or has just discovered a new website he starts engaging with, which reduces his time on Stack Overflow. We also expect these effects to average out over the 174 legendary users.
- Finally, we believe that focusing on Legendary users restrains the analysis to extremely active users which is not necessarily the most interesting user group to study for actionable social science findings. Findings showing how to motivate passive users to contribute to the creation of this common knowledge base would certainly be much more valuable to a Stack Overflow decision maker since passive users are logically the ones one would primarily want to change the behavior of.

b. Next steps

Our envisioned next steps are:

- Analyze the statistical significance of our findings by calculating p-values (and check whether they are < 5%) as well as calculating the confidence intervals of the relative changes in each of the interrupted time-series designs
- Perform additional analysis on the same data and in particular: an interrupted time-series design on the average gain of reputation per day for Legendary users (before and after getting the badge). This would allow us to look at more granular behavior: maybe, just after getting the badge, the frequency of great days decreases but the daily reputation gain remains constant.
- Replicate the entire analysis we have performed with other badges – especially lower status badges which would allow us to study even more interesting groups of users (in terms of potentially actionable findings).

Conclusion:

The question of the motivations behind mutual help on online social Q&A platforms is vast and complex. By describing the behavior of very active users on Stack Overflow before and after being rewarded with a badge, we intended to provide more context and understanding of the impact of gamification on social Q&A platforms' users.

Through our quest to show if and how gamification impacts user behavior on Stack Overflow, we found interesting results. Focusing on Legendary users – the ones who got a daily reputation score higher than 200, 150 distinct times - we performed an interrupted time-series analysis, which showed that Legendary users change behavior after getting their Legendary badge. Indeed, the slopes of the linear approximations of the average number of posts per day are different in both regimes: from a positive slope showing an increasing engagement just before getting the badge, to a negative slope showing a decreasing engagement after getting the badge. This result was obtained by comparing periods of 30 days before and after getting the badge, but the finding also held for 300 days before and after getting the badge. This suggests gamification actually has a long-term impact on Legendary users, impacting their engagement in a time window of at least 10 months before and after they get their badge. The interrupted time-series design applied to the frequency of "great days" before and after getting the Legendary badge shows the frequency of "great days" also decreases after badge date but starts increasing again after 10 days after badge date, testifying to a rise in up votes of these users' posts, starting 10 days after they became Legendary.

Bibliography:

Causal Discovery in Social Media Using Quasi-Experimental Designs, Hu̇seyin Oktay, Brian J. Taylor, David D. Jensen, University of Massachusetts Amherst, 2010

How Social Q&A Sites are Changing Knowledge Sharing in Open Source Software Communities, Bogdan Vasilescu, Alexander Serebrenik, Premkumar Devanbu, Vladimir Filkov, Eindhoven University of Technology and University of California Davis, 2014

Self-Interest, Reciprocity, and Participation in Online Reputation Systems, Chrysanthos Dellarocas, Ming Fan, Charles A. Wood, MIT Sloan School of Management, 2004