Quick Trigger on Stack Overflow: A Study of Gamification-influenced Member Tendencies

Yong Jin*, Xin Yang*, Raula Gaikovina Kula†, Eunjong Choi†, Katsuro Inoue†, Hajimu Iida*
* Nara Institute of Science and Technology, Japan
{jin.yong.jq0, kin-y}@is.naist.jp, iida@itc.naist.jp
† Osaka University, Japan
{raula-k, choi-e, inoue}@ist.osaka-u.ac.jp

***Abstract*—In recent times, gamification has become a popular technique to aid online communities stimulate active member participation. Gamification promotes a reward-driven approach, usually measured by response-time. Possible concerns of gamification could a trade-off between speedy over quality responses. Conversely, bias toward easier question selection for maximum reward may exist. In this study, we analyze the distribution gamification-influenced tendencies on the Q&A Stack Overflow online community. In addition, we define some gamification-influenced metrics related to response time to a question post. We carried experiments of a four-month period analyzing 101,291 members posts. Over this period, we determined a Rapid Response time of 327 seconds (5.45 minutes). Key findings suggest that around 92% of SO members have fewer rapid responses that non-rapid responses. Accepted answers have no clear relationship with rapid responses. However, we did find that rapid responses significantly contain tags that did not follow their usual tagging tendencies.**

I. Introduction

With the prominent use of libraries in everyday software systems, system maintainers and developers alike, face difficulties in understanding either a new or familiar but improved version of a library. [1]. In recent times, developers have turned to online forum communities like Stack Overflow (SO)[1] for learning resources, most commonly in the form of either code examples or advice from the online community. Studies show evidence of effectiveness [2], [3].

Like most online forums, to stimulate member activity and retention SO employs reward-driven gamification techniques depicted by member reputation and status. Members gain reputation by contribution and evaluation scores (voting) by their peers.

According to the SO documentation[2], reputation is described as a rough measurement of how much the community trusts the member, and how knowledgeable the member is. The primary way to gain reputation is by posting good questions (as a questioner) and useful answers (as an answerer). Members can earn a maximum of 200 reputation per day from any combination of the activities, such as Bounty awards and accepted answers. An *accepted answer* (AA) will gain 15 reputation (+2 to the accepter), which is the highest and quickest way to earn reputation rewards. Every question post contains tags[3], meant to classify the post content. One question could contain several tags related to various knowledge areas. A SO member's profile contains their tag information that is ranked by the a members usage.

TABLE I
SO Information Sources

	SO XML Table	Metrics
Creation Date	Post	RT, RR and RT Threshold
Accepted Answer	Post	Accepted Answer AA
Member Tags	User, PostTags, Posts	Tag Score (TS)

Gamification influences the time to respond to a question. The quickest response, defined in this paper as *Rapid Response* (RR) often 'wins' the most reward. This 'quick trigger' effect, can be viewed as being beneficial to the livelihood of the SO community, the questioner and members as it promotes active feedback. However, the need to 'win' rewards may raise concerns related to post quality and 'true' forum knowledge. Research today has turned to SO as an information source to mine knowledge and expertise on various topics [4], [5].

In this study, we focus not necessarily on the post context quality but to what extent gamification-techniques have influenced member response tendencies. The following questions guide the study:

- RQ1: Is Rapid Response (RR) widespread among community members?
- RQ2: Are most Accepted Answers (AA) related to RR?
- RQ3: Does member tagging tendencies change with RR?

Our approach is to mine and analyze SO information sources to understand how gamification has influenced member tendencies.We used the SO data provided by Mining Software Repositories 2015 mining challenge [6].

II. Gamification-influenced metrics

Table I details the information sources used to define our gamification-influenced metrics. We used the structure from the available public data dump [4] for our metric attributes.

[1]http://stackoverflow.com
[2]http://stackoverflow.com/help/whats-reputation
[3]http://stackoverflow.com/help/tagging
[4]http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

978-0-7695-5594-2/15 $31.00 © 2015 IEEE
DOI 10.1109/MSR.2015.57

A. Rapid Response (RR)

Suppose post p is identified as either a question $p(q)$ or answer $p(a)$. Let $createTime$ be the creation time of a post. Hence, Response Time (RT) measured in the unit of seconds:

$$RT(x) = x.createTime - y.createTime \\ \rightarrow (\exists y(x.parentId = y.id) \wedge x = p(a) \wedge y = p(q)) \quad (1)$$

where x is an answer post, y is a question and $p(q).Id$ and $p(a).parentId$ match. We define *Rapid Response* (RR) as responses intended by response time to 'win' reputation points. We determine rapid response by using a threshold of fast responses. Therefore, for a given threshold, Rapid Response (RR) are the posts that equal or less than a given RT threshold. Non-Rapid Response (NRR) are the ones greater than the RT threshold. Hence:

$$RR = \{p|RT(p) \leq RTthreshold\} \quad (2)$$

and correspondingly,

$$NRR = \{p|RT(p) > RTthreshold\} \quad (3)$$

In order to identify extreme response times, we apply the *pareto principle* as the $RTthreshold$. In this case, we believe that 20% of the RT represent around 80% of all responses. We have used such thresholds in previous works[7].

B. Accepted Answer (AA)

An accepted answer is an attribute of the post table and it marked for every question in $p.AcceptedAnswerId$ which corresponds to $p(a).Id$. We assume that for post p, that $p \in RR$ and $p \in AA$ may be an indication of a gamification-influenced to quickly 'win' reputation. Conversely, posts not accepted as answers are referred to as Non-Accepted Answers (NAA).

C. Tag Score(TS)

Tag score allows us to measure tendencies of member tagging. We use a normalized metric of tag score (TS) based on the frequency of use. Suppose a post p contains unique tags t where $p = \{t_1, t_2, t_3, ..., t_n\}$. For a member m with the unique $m.id$, the tag score(TS) for a tag is:

$$TS(t_x, m) = \frac{|t_x \in p|}{|p|} \rightarrow (m.id = p.OwnerUserId) \quad (4)$$

TS ranges from 0-1 (1 indicates tags are consistent). Lower scores indicate that this tag is not frequently used by that particular member. For example, suppose a member m has 5 posts with the following tags, $p1 = \{\texttt{\#java}, \texttt{\#sql}\}$, $p2 = \{\texttt{\#java}\}$, $p3 = \{\texttt{\#java}, \texttt{\#sql}\}$, $p4 = \{\texttt{\#java}\}$ and $p5 = \{\texttt{\#java}, \texttt{\#sql}, \texttt{\#javascript}\}$. All 5 posts contained tag `#java`, 3 contained `#sql` and 1 has the `#javaScript` tags respectively. Therefore, the TS for each tag is: `#java` = 5/5 = 1, `#sql` = 3/5 =0.6 and 1/5 = 0.2 for `#javascript` respectively. Therefore, ts scores for all posts are: $p1 = \{1, 0.6\}$, $p2 = \{1\}$, $p3 = \{1, 0.6\}$, $p4 = \{1\}$ and $p5 = \{1, 0.6, 0.2\}$

For our analysis, we compare the TS between RR and NRR posts. Our assumption is that gamification causes members to explore tags outside of their usual, just to gain quick reputation 'wins'.

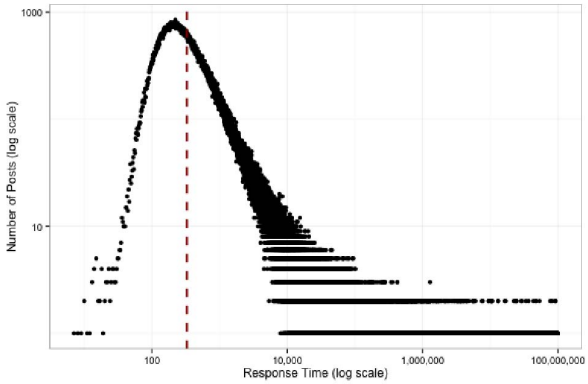

Fig. 1. Distribution of Rapid Responses

III. Experiment and Results

A. Experiments

To test our approach, we conducted an analysis on a sample representative dataset. As shown in Table II, we mined four months of answer posts $p(a)$ which spanned from June 1, 2011 to September 30, 2011. Then, we recovered their corresponding $p(q)$ to calculate RT. From 1,223,088 posts, we retrieved 795,667 answer posts from 101,291 SO members. It includes 159,575 rapid responded answer posts, 636,036 non-rapid responded answer posts. There were 56 instances that we could not calculate the response time due to the answer time was earlier than question time, which we treat as noisy data.

B. Rapid Responses

Table IV details the RT and RR threshold calculated for the dataset. As shown, the RT threshold = 327 seconds or 5.45 minutes (20% of the RT distribution). This distribution, shown

TABLE II
Experimental Dataset

	SO Dataset (4 months)
time period	2011-06-01 to 2011-09-30
# of members	101,291
# of posts (p)	1,223,088
# of question posts (p(q))	446,647
# of answer posts (p(a))	795,667

TABLE III
SUMMARY OF DISTRIBUTIONS FOR RT, RR ONLY MEMBERS, NRR ONLY MEMBERS, AND TAGS

	RT in seconds All members	RT in seconds RR Only (NRR = 0)	NRR Only (RR = 0)	Tags Usage Count by all members
# of Members	101,291	1,700 (1.7%)	81,849(80.8%)	101,291
Min	7	18	328 (5.5 mins)	1
1st Qu.	399 (6.6 mins)	178 (2.9 mins)	3,096 (51.6 mins)	3
Median	1,291 (21.5 mins)	231 (3.8 mins)	55,682 (15.5 hours)	5
Mean	2,833,432 (1.1 months)	223 (3.7 mins)	8,323,968 (3.2 months)	12
3rd Qu.	17,124 (4.7 hours)	276 (4.6 mins)	3,797,704 (1.5 months)	10
Max	99,325,885 (3.1 years)	327 (5.4 mins)	99,325,885 (3.1 years)	1,137

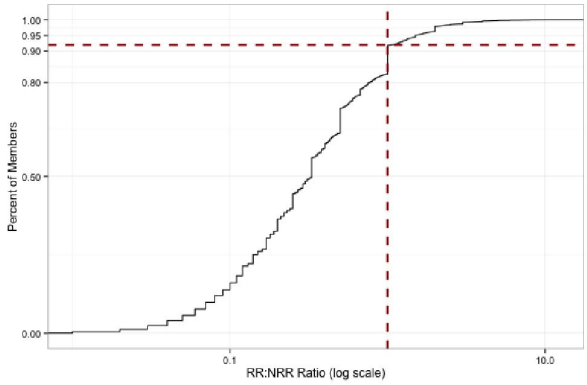

Fig. 2. RR:NRR Ratio Distribution among members

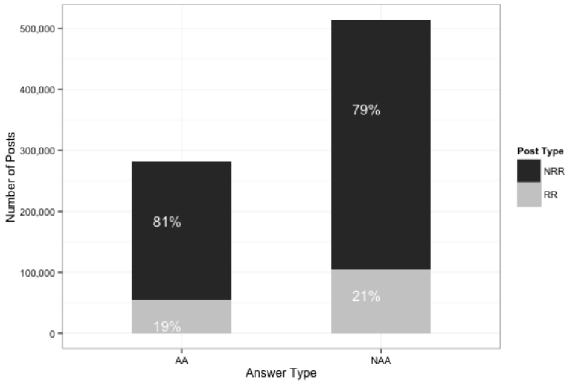

Fig. 3. Analysis of AA, NAA, RR and NRR combinations

in Figure 1 further supports our pareto principle that 20% of RT represents about 80% of the posts. From Table III, we see that the faster response is 7 seconds and longest detected was 3.1 years response-time. Based on all the answer posts of four months, we found that there are many question posts are earlier than this period. In this study, we also include these question posts in dataset.

To find the distribution tendencies for all members, we calculated a ratio of RR:NRR. Table III shows the distribution of members that either had only RR Only (1.7%) or NRR Only (80.8%). These are members that only have RR or NRR respective posts. As for the rest, we find that 159,575 (20.1%) are RR while 636,036 (79.9%) are NRR. Figure 2 depicts that almost 92% of members have more NRR than RR posts (where RR $<$ NRR). Based on these results, to answer RQ1, *the ratio of RR is still lower that NRR posts for SO members, so not as widespread among the community*.

TABLE IV
RAPID RESPONSE AND ACCEPTED ANSWER STATISTICS

	Experimental Data
RT Threshold	327 (seconds)
# RR	159,575
# NRR	636,036
# AA	281,110
# NAA	514,557
# Single Answer	248,224

C. Accepted Answers

As shown in Table IV, it is interesting to note that 248,224 question posts (around 56% in all question posts) have only one answer in the experimental dataset. In these sole-answered question posts, 134,204 question posts (around 54%) have AA, leaving 114,020 (around 46%) have no AA (i.e. NAA). The remaining 198,423 question posts (around 44%) which have more than one answer, 146,906 question posts (around 74%) have AA, leaving 51,517 (around 26%) have no AA. These results suggest that *questioners tend more to accept answers from a list of multiple answers (74%), instead of single answers (54%).*

As depicted in Figure 3, we can observe no direct relationship between RR and AA. With all the answer posts, 281,110 (around 35%) are AA, and 514,557 (around 65%) are NAA. The results show that 19% of AA are also RR. NAA that are RR have similar proportions(around 21%). Based on these results, to answer RQ2, *there seems to be no direct relation of AA in relation to RR.*

D. Tagging Tendencies

Table III shows that the for each member the median tag usage is 5 tags and 12 on average. Also, the most unique tags used by one used member is 1,137 tags. Figure 4 depict a comparison of tag scores (ts) between RR and NRR posts. The results suggest that that RR posts contain tags of a lower score than NRR posts. This difference was found to be

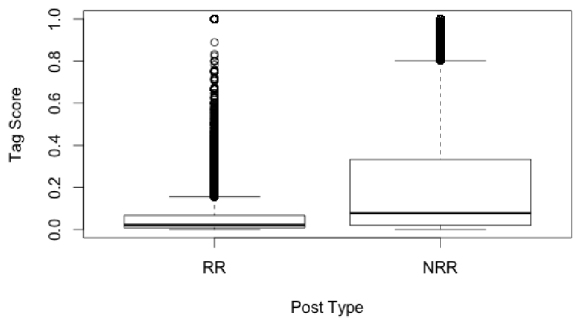

Fig. 4. Distribution of Tag Score in RR and NRR Posts

statistically significant (p-value$<$0.05) by the Mann-Whitney-Wilcoxon test. This result suggests that RR answer posts tags were more dispersed than NRR answer posts. It implies that overall for RR posts, SO members tend to use tags that are not as consistent compared to their usual responses. So to answer RQ3, *member tagging tendencies does change with RR posts.*

IV. Discussion

A. Implications

Member tendencies describe the norm or prevailing disposition within that community. As gamification becomes a popular tool to engage community members, this study could be a start of investigations on how gamification techniques can affect these natural tendencies to the point of affecting quality. For instance, RR could be considered 'quick and dirty' or may cause members to 'avoid' legitimate questions that are not worth their effort.

The results indicate that RR, which is a consequence of gamification, significantly affects the tagging tendencies of members. However, overall the accepted answers and members inclination to rapid response is not as affected. For future work, we would like to expand to other gamification-influenced metrics covering more the context of the post such as the post body and headings. Other topics of interest would be questions that RR avoid or attract. Additionally, other aspects of reputation scoring such as voting may be of importance.

B. Threats to Validity

A threat to validity is the use of the pareto principle as the RT threshold measure. Sinha *et al.* have examined the activeness of Stack Exchange users and they found the large amount of activities is done by a small group of people that satisfies 80-20 rule [8]. Other outlier identification methods could be used such as Tukeys Outlier Filter [9]. We envision this for future work and are confident as the Figure 1 validates our use. Another threat is size of the dataset. We feel that a coverage of over 1 million posts is sufficient for this preliminary experiments. Future work will include looking at datasets that cover a longer time period.

V. Related Work

Our work is complementary to the following: Bhat *et al.* have found the tag-related factors have a strong relationship with the response time of question [10]. Wang has performed an analysis to identify different life-cycle patterns of questions in order to investigate how to make a question be answered faster [11]. Bosu et al. analyzed the dynamics of reputation in SO and their result can be used to guide users to gain reputation faster [12]. Bazelli *et al.* investigate the personality traits of SO users, and the results show that top reputed contributors are more extroverted and have less negative emotion [13].

VI. Summary

We present a study of how gamification affects online community members tendencies in terms of response time. Results indicate that most members do not undertake in such rapid response (RR) activities. However, we found when they did provide RR, the tags did not match their usual tagging tendencies. We would like to investigate the rapid responses contexts and compare with other sources of reputation such as earning votes on posts in future work.

References

[1] S. Subramanian, L. Inozemtseva, and R. Holmes, "Live api documentation," in *Proc. of Internl Conf. on Soft. Eng.*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 643–652.

[2] P. C. Rigby and M. P. Robillard, "Discovering essential code elements in informal documentation," in *Proc. of Internl Conf. on Soft. Eng.*, ser. ICSE '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 832–841.

[3] M. Linares-Vásquez, G. Bavota, M. Di Penta, R. Oliveto, and D. Poshyvanyk, "How do api changes trigger stack overflow discussions? a study on the android sdk," in *Proc. of Internl Conf. on Prog. Comp.*, ser. ICPC 2014. New York, NY, USA: ACM, 2014, pp. 83–94.

[4] A. Barua, S. Thomas, and A. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," *Empirical Software Engineering*, vol. 19, no. 3, pp. 619–654, 2014.

[5] E. Choi, R. G. Kula, N. Yoshida, and K. Inoue, "What do practitioners ask about code clone? a preliminary investigation of stack overflow." in *9th Internl Work. on Soft. Clones, IWSC 2015, Montreal, Canada, March 6, 2015*, 2015.

[6] A. T. T. Ying, "Mining challenge 2015: Comparing and combining different information sources on the stack overflow data set," in *The 12th Work. Conf. on Min. Soft. Repo.*, 2015, p. to appear.

[7] R. G. Kula, A. Cruz, N. Yoshida, K. Hamasaki, K. Fujiwara, X. Yang, and H. Iida, "Using profiling metrics to categorise peer review types in the android project," in *Soft. Rel. Eng. Work.(ISSREW), 2012 d Internl Symp.*, Nov 2012, pp. 146–151.

[8] V. S. Sinha, S. Mani, and M. Gupta, "Exploring activeness of users in qa forums," in *Proc. of Work. Conf. on Min. Soft. Repo.*, ser. MSR '13, 2013, pp. 77–80.

[9] J. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[10] V. Bhat, A. Gokhale, R. Jadhav, J. Pudipeddi, and L. Akoglu, "Min(e)d your tags: Analysis of question response time in stackoverflow," in *Adv. in Soc. Net. Anal. and Min. (ASONAM), Internl Conf.*, Aug 2014, pp. 328–335.

[11] Y. Wang, "Making your programming questions be answered quickly: A content oriented study to technical q a forum," in *Collab. Comp.: Net., Apps. and Workshar.(CollaborateCom), Internl Conf.*, Oct 2014, pp. 368–377.

[12] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft, "Building reputation in stackoverflow: An empirical investigation," in *Proc. of the Work. Conf. on Min. Soft. Repo.*, ser. MSR '13, 2013, pp. 89–92.

[13] B. Bazelli, A. Hindle, and E. Stroulia, "On the personality traits of stackoverflow users," in *Proc. of Soft. Main. (ICSM), Internl Conf.*, Sept 2013, pp. 460–463.