One-day flies on StackOverflow

Why the vast majority of StackOverflow users only posts once

Rogier Slag,* Mike de Waard,† Alberto Bacchelli‡
{*r.g.j.slag, †m.dewaard}@student.tudelft.nl, ‡a.bacchelli@tudelft.nl
Delft University of Technology, The Netherlands

***Abstract*—StackOverflow (SO) is a popular question and answers (Q&A) platform related to software development. An interesting characteristic of SO is that about half of its users makes only one contribution to the platform in total.**

In this work, we study this group of users, which we call *one-day flies*, and we investigate why they do not continue to contribute to the platform, despite its growing popularity. By achieving this understanding we can find ways to enable users to become more active, thus improving the platform.

I. Introduction

StackOverflow is a Q&A platform and community on the Internet that focuses primarily on programming related questions. SO has been gaining huge popularity during the last few years [1] and the availability of its data dump to the public has triggered numerous investigations (*e.g.*, [10], [3], [7]). Many studies on SO cover aspects related to users, such as their personalities [5], their field of expertise, and on how they contribute [8]. Little is know, yet, about the underlying structure of the StackOverflow community, in particular the distribution of posts and reputation among users.

In this paper, we try to start filling this gap by analyzing less active users, *i.e.*, those who post only once during their whole registration period. To this aim, we first quantify these users finding that they account for about half of the users of SO. Then, we use different information sources to try to quantitatively characterize this set of users, which we call *one-day flies*, by contrasting them with more active participants. Finally, we conduct an initial qualitative investigation to better understand what could induce these users to stop contributing. Results show that many of our hypotheses of why users become *one-day flies* do not fully explain this phenomenon. We discuss aspects emerged as candidates for future evaluation.

II. Related work

Several aspects of SO have already been investigated, such as how to build reputation [5] and how the reputation system encourages users to participate [8]. These studies gave us insight in how the reputation system works and how this affects the user base. Correa & Sureka [6] investigated the deletion of questions on SO; we use this as the basis for verifying some of our hypotheses. Bazelli *et al.* analyzed personality traits of SO users [4]; on their work we base our sentimental analysis for post and answers. Yang *et al.* investigated the group of experts on SO [11]; our work complements theirs by considering a different group of users.

III. Defining one-day flies

As shown by several studies [11], Q&A platforms are composed by users with different activity behaviors, which closely follow a power law distribution: A small set of very highly active users contributes to the vast majority of the produced content; a larger set of less active users contributes to fewer parts of the content; and a set (the largest) of users contributes very little to nothing.

We start our investigation by quantifying the result of this community dynamics in SO and use it to cluster users and define *one-day flies*. We consider the SO data dump provided for the MSR Challenge 2015 [12] and we analyze both user activity (based on number of posts) and user reputation (based on the score obtained in the gamification mechanism in place on SO [2]). We find that 47% of the users made a maximum of one post, while the most prolific user had 30.043 posts. Similarly, 60% (2.063.174) of the users had a reputation of one, the third quartile is 13, and the maximum value is 709.269. This is in line with general trends on Q&A platforms.

We then apply clustering to users based on their reputation, using k-means. The largest group consists of 3.469.897 users, who have between 0 and 1 of reputation and posted questions; the medium active user group is significantly smaller (3.012 users) with an average reputation of 38.000; the group with the most reputation is by far the smallest (186 users) with a very high average reputation of ca. 200.000.

With this information, we define *one-day flies* as users who posted a maximum of one question in SO, thus covering both low activity and low reputation. We also consider users that are registered for more than 6 months to avoid that recently registered users could bias the results. This set comprises 1.622.688 users, *i.e.*, 47% of the user base.

IV. Experiment

We conduct an experiment to quantitatively test different hypotheses, expressed as research questions on what may characterize *one-day flies*. In the experiment we use the complete SO MSR Challenge dump [12], imported into a MSSQL database. For the actual querying, statistical analysis, and plotting of the data we use a mixture of Ruby (especially to mine additional data), C# (to query the data), and R (to perform statistical analysis and generate plots). In the following we describe each hypothesis, the intuition behind it, the research method we used to test it, and the results.

RQ1: Duplicate Questions

Are *one-day flies* more likely to create questions eventually marked as duplicates?

Intuition. New users who register on the website could, because of their unfamiliarity with the system, ask questions that have been asked previously. These questions would be flagged by the community as *duplicate*. Although this may be the appropriate action for the community as a whole, it could be extremely discouraging for the user which posted the question, thus inducing to not post anymore.

Research method. To answer this research question, we compare the percentage of duplicate questions between and *one-day flies* and the other users (defined from now on as *regular users*). To do so, we extract questions asked by *one-day flies* and count those marked as duplicate by checking on LinkTypeId in the PostLinks table; then we do the same for questions asked by regular users. After retrieving these results we compare them.

Results. The total amount of questions asked by *one-day flies* is about 8%. The number of duplicate questions is surprisingly low for both groups. One could expect that this would be high on a public Q&A community where a large percentage of the users is unfamiliar with platform. However, only 2.2% of the one-day flies questions is marked as a duplicate. Surprisingly, this percentage increases to 2.9% for regular users; we expected regular users to be more familiar with the platform and its search functionality. Based on these results, *one-day flies* are not more likely to create duplicate questions than regular users, but slightly less.

RQ2: Uncommon Tags

Are *one-day flies* more likely to post questions with uncommon tags?

Intuition. The topic of a question in SO is commonly translated into tags, which make it easier for experts to find questions they could answer. Users who post questions regarding a topic with hardly any experts on the SO platform can get discouraged by the lack of comments or answers to guide them into the right direction. This types of topics would be reflected by uncommon tags accompanying the questions.

Research method. Correa and Sureka found over 3.300 tags that only occur in questions of quality far below StackOverflow standards [6]. The answer to our research question is inspired by their findings, as similar results might be found. We analyze the top 50 most popular tags given by *one-day flies* questions in comparison to the top 50 tags given in all questions in SO. This to get insights on the most popular tags and any popular uncommon topics among *one-day flies*.

Results. Surprisingly the top-5 tags of both *one-day flies* and regular users are identical (although ordered different). Although there are a few tags that are specific to *one-day flies*, most of them is very similar. Based on these results, the answer to our research question is negative.

RQ3: Deleted Questions

Are *one-day flies* more likely to have their posts removed (either by themselves or by a moderator)?

Intuition. Another possible explanation for the large number of users with a low question count can also be found in deletions. Users themselves can delete questions after posting them, and so can the moderators in the community. This can have quite an impact. As studied in [6], deleted questions usually are of a very low quality (the quality is considered low compared to the expectation of the SO community). Users unfamiliar with the ethos of StackOverflow may inadvertently post such questions, only later realizing these do not belong there, and be deleted (or be deleted by moderators).

Research method. The number of deletions is not directly obtainable from the database dump as provided by StackOverflow: The only way to know whether a question on the data dump has been eventually removed is to check its online existence on the SO website. Due to our resources we could not check each question in the dump against the website, so we use random sampling. We select two samples of 150.000 questions: One for *one-day flies* and one for people with a high post count. For each of these sets, we got the SO page and parsed its DOM structure. This allowed us to extract whether the question was deleted, or possibly closed.

Results. The number of deleted questions for *one-day flies* was 15.4%, whereas it was 10.9% for the high post count users. However, it cannot easily be determined what the reason for deletion was. As explained in [6], there are several reasons for a question to be deleted. Therefore further investigation would be necessary to give a more complete answer to this question. The same analysis also gave interesting insights in the number of closed posts. For both groups, the ratio of closed posts is surprisingly low: 0.92% for *one-day flies* and 0.76% for high post count users. Questions can be closed for various reasons, such as being a duplicate or for being off-topic. Closed questions may be improved or deleted later on. The difference is low, thus we can give a negative answer to the research question when considering closed posts.

RQ4: View Count

Are *one-day flies* more likely to attract less views?

Intuition. Users' questions may be not viewed as much as they had anticipated, resulting in fewer good answers. It may be the case their question was listed lower due to their lack of reputation, or simply got lost in the list of questions. It is also possible that questions from people with no post history simply get opened less often, resulting in lower quality answers. A low view count can discourage the question askers, since they may feel neglected by the community.

Research method. To answer our research question, we distinguish between *one-day flies* and all users. To examine the results, we eliminate the highest viewed questions and the least viewed questions from the data set: We consider different cut-offs (*i.e.*, 5%, 10%, and 20%) and compute separate results.

Results. Both groups have an average view count of ca. 190 (198 for *one-day flies*, and 187 for regular users), but the standard deviation is high (400 for *one-day flies*, over 1.000 for all). To compensate for these effects, the dataset was modified to only include the posts which fitted into the 10% to 90% range of the view counts and the results were rerun. Using this reduced data set, with the extremes removed, we obtain a similar result, with an average view count of 127 for *one-day flies* and 113 for all. The standard deviations dropped a lot as well, to only 100 respectively 93. When the analysis is repeated by eliminating the extreme 5% of 20%, similar results remain. Therefore we conclude that the posts of *one-day flies* are actually not viewed less.

RQ5: Unanswered Questions

Are *one-day flies* more likely to receive no answer to their questions?

Intuition. New users simply got let down by the community: They asked a question which did not violate any of the StackOverflow rules, but did not receive a (satisfactory) answer anyway. This would probably the most discouraging situation, since the user did not receive help due to aspects out of his or her control. That might cause them to lose the confidence in the community itself.

Research method. To answer our question, we compared the set of questions done by *one-day flies* to the set of questions done by other users.

Results. Based on the data, we found the first non-negligible difference between the groups. *one-day flies* have a larger percentage of unanswered questions (*i.e.*, 17%) compared to the other group (*i.e.*, 10%). It could be that our intuition is correct and part of the *one-day flies* leave SO because they get demotivated by their questions left unanswered. Especially if they saw other users receive answers to their posts much more often. Since this difference is only 7% we can answer the our research question positively, but it is sufficient to account for the vast number of *one-day flies*.

V. Manual investigation

Since the results did not reveal a clear reason that could completely account for the whole large number of *one-day flies*, we decided upon a small qualitative research to get new ideas and see if something was overlooked that can be found in the data. To this aim, we manually analyzed 50 posts and their accepted answer. We found the following patterns: (1) 20 questions and 19 accepted answers contained code; (2) 48 answerers were friendly, 2 were unfriendly; and (3) 11 times the answer was easily found by a search engine, the other 39 times they tried to find a solution themselves.

Pattern 1 shows that over half the questions and accepted answers did not contain code, even though the questions where programming related. Most of them where regarding versioning systems, or development methods (such as Agile or database systems). To get better insight on this pattern, we did an automated search and found that for non *one-day flies* 65% of the questions had code and 58% of the answers (over the set of all questions with an accepted answer). For *one-day flies* these numbers were 70% and 65%, respectively. This indicates *one-day flies* give and receive a bit more code than the other users. During our manual analysis we found the code example to be of rather poor quality, we did not investigate this further, but future work could consider this aspect.

We found pattern 2 is interesting: Two answers where unfriendly, but still got accepted. We investigated this further by using a Ruby script [9] to evaluate how friendly the answers and questions were on a scale of -4 (very unfriendly) to 4 (very friendly) on a larger (10.000) subset of questions. Results showed that non *one-day flies* questions had a niceness of 0.079, and the answers they received had a niceness of 0.059. This indicates that both the questions and answers were mostly formal and professional, and on average slightly positive. For *one-day flies*, this was again a bit different: They had a question niceness of 0.196 and received an answer with niceness 0.092. Still the tone appears to be mostly formal and professional, but the underlying touch is a bit more positive.

Pattern 3 gives the idea that askers are relatively "lazy." However, since the classification "lazy" is quite opinionated, this cannot be used as a direct research result. However we found that a few questions where later answered by the question owner him/herself. This gave the idea that one-day flies might only use SO for self-documentation purposes.

VI. Discussion

The presented results are not conclusive about what could motivate the behavior of *one-day flies*. As indicated by the previous sections, there does not seem to be a clear reason within these statistics why *one-day flies* contribute less to the platform. However, our results gave insights on some aspects that might be interesting venues for future work, we describe some of those in the following.

A. Code example quality

In general it seems the quality of the code examples in questions of *one-day flies* is lower compared to the quality of users with a higher reputation. *One-day flies* generally include too much code (by simply pasting all the code into the questions code block). This discourages users to form an adequate explanation.

Apart from that, the general code quality itself is lower as well. Code is poorly modularized, which causes people to respond to those issues, before addressing the original problem of the question asker. He or she may then get irritated, since they receive feedback they simply are not ready for at that point in the learning process. The same goes for the people trying to figure out what the issue is: They need to wrap their mind around to some non-standard principles used by others, which are (from their perspective) less easy to reason about.

B. Negative feedback

As stated in the previous paragraph, it is often the case the StackOverflow community gives tips and advice to novices

how certain problems should be structured. However, for the question asker this can be seen as negative feedback. It does not solve their problem, while it increases the work they have to do to solve the problem or even get community support.

Even without this feedback, the StackOverflow community has strict rules on the type of questions allowed on the platform. A quick search over time showed the authors that these rules were not strictly followed in the early days of the StackExchange platform, but are more followed these days. This causes many beginners questions, subjective discussions (*Is A better than B?*), or too-localized questions (*What does this regular expression do?*) to be voted for closing or deletion.

Since the community has answered a large number of questions, one can expect the same thing to happen for other things, such as *marked as duplicate*. The chances are increasing that somewhere in the database, a very similar question might have been asked already, and received an appropriate answer. Meanwhile, a beginner on the platform might not be aware of the exact terminology to find that specific question. Therefore they may consider their problem to be new and unique, causing them to post it. Several minutes later, they may find their question marked as a duplicate combined with a snide comment on "how you can search on StackOverflow." It is no wonder that a user could see this as a negative feedback from the community.

C. Self-answering

In section V, we found that some of the questions accepted answers were given by the askers themselves. This could be done for documentation purposes, but also for attempting to gain reputation, while other answers might be better. However in order to explore this further, it is first interesting to evaluate whether *one-day flies* answer their own questions more often than regular users.

A user does not get reputation for answering their own question (but they can get reputation if the answer is upvoted by others). Here there seems to be a clear trend: users have found and answer, and decided to share it with the community (even though they do not have a direct profit from it themselves). This seems to be an indication these users are not scared away from StackOverflow at that point (otherwise they likely would not have posted that answer).

D. Everything can be found

During the research we had several discussions with software developers regarding this high amount of *one-day flies*. In most of these sessions developers said to never ask a question on StackOverflow because the answer to their question was already available. This gave the authors a new idea that possibly almost all questions can be found, thus the chance that a user has two questions that are not answered on StackOverflow already is rather slim. This could cause the community to look like it has many *one-day flies* whereas the *one-day flies* have only one question that could not be found on StackOverflow before. If this is the case, then when taking into account the user growth, one should see a decrease in the amount of questions asked over time.

VII. Conclusion

In this paper we analyzed an aspect of the underlying structure of the SO community: the behavior of *one-day flies*. These are users who post only one question during their participation to SO, even if they have been registered for six months at least.

We investigate five research questions, based on different information sources, to verify intuitions we have on the behavior of *one-day flies*. Results show that *one-day flies* are not more likely to create duplicate questions, to use uncommon tags, or to receive less views for their questions. But their questions do get removed slightly more frequently, either by themselves or a moderator, and they are more likely (7%) to receive no answer to their question. Although these two facts could explain part of the reason why *one-day flies* participate less, they do not seem to account for the whole story. More research is necessary to understand better this phenomenon.

As a first step in this direction, we did a manual investigation and suggested future research directions.

References

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

[2] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, pages 95–106. International World Wide Web Conferences Steering Committee, 2013.

[3] A. Barua, S. W. Thomas, and A. E. Hassan. What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical Software Engineering*, 19(3):619–654, 2014.

[4] B. Bazelli, A. Hindle, and E. Stroulia. On the personality traits of stackoverflow users. In *Software Maintenance (ICSM), 2013 29th IEEE International Conference on*, pages 460–463. IEEE, 2013.

[5] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft. Building reputation in stackoverflow: an empirical investigation. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, pages 89–92. IEEE Press, 2013.

[6] D. Correa and A. Sureka. Chaff from the wheat: characterization and modeling of deleted questions on stack overflow. In *Proceedings of the 23rd international conference on World wide web*, pages 631–642, 2014.

[7] P. Morrison and E. Murphy-Hill. Is programming knowledge related to age? an exploration of stack overflow. In *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on*, pages 69–72. IEEE, 2013.

[8] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos. Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 886–893. ACM, 2013.

[9] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[10] C. Treude, O. Barzilay, and M.-A. Storey. How do programmers ask and answer questions on the web?: Nier track. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 804–807. IEEE, 2011.

[11] J. Yang, K. Tao, A. Bozzon, and G.-J. Houben. Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In *User Modeling, Adaptation, and Personalization*, pages 266–277. Springer, 2014.

[12] A. T. T. Ying. Mining challenge 2015: Comparing and combining different information sources on the stack overflow data set. In *Proceedings of MSR 2015*, page to appear, 2015.