Journal of
the Brazilian Computer Society
a SpringerOpen Journal

RESEARCH Open Access

A case study of contributor behavior in Q&A site and tags: the importance of prominent profiles in community productivity

Adabriand Furtado[*], Nigini Oliveira and Nazareno Andrade

Abstract

Background: Question-and-answer (Q&A) sites have shown to be a valuable resource for helping people to solve their everyday problems. These sites currently enable a large number of contributors to exchange expertise by different ways (creating questions, answers or comments, and voting in these), and it is noticeable that they contribute in diverse amounts and create content of varying quality.

Methods: Concerned with diversity of behaviors, this paper advances present knowledge about Q&A sites by performing a cluster analysis with a multifaceted view of contributors that account for their motivations and abilities to identify the most common behavioral profiles in these sites.

Results: By examining all contributors' activity from a large site named Super User, we unveil nine behavioral profiles that group users according to the quality and quantity of their contributions. Based on these profiles, we analyze the community composition and the importance of each profile in the site's productivity. Moreover, we also investigate seven tag communities from Super User aiming to experiment with the generality of our results. In this context, the same nine profiles were found, and it was also observed that there is a remarkable similarity between the composition and productivity of the communities defined by the seven tags and the site itself.

Conclusions: The profiles uncovered enhance the overall understanding of how Q&A sites work and knowing these profiles can support the site's management. Furthermore, an analysis of particularities in the tag communities comparison relates the variation in behavior to the typical behavior of each tag community studied, what also draws implications for creating administrative strategies.

Keywords: Q&A sites; Empirical methods; Quantitative; Datamining and machine learning; Studies of Wikipedia web

Background

Question-and-answer (Q&A) sites currently enable thousands of people to help each other to find problem solutions. In these sites, users exchange their expertise through activities like posting questions, answers or comments, and voting on the quality of these posts, in which it is useful to identify worthy content. Sites such as Yahoo! Answers and StackOverflow have shown the potential to leverage massive numbers of voluntary contributors to create valuable and dynamic knowledge bases [1-3].

As a result of a sizeable community, it is expected that contributors from Q&A sites exhibit diverse behavior in creating content. For example, users have different motivations to contribute [3-5], spend varying amounts of time contributing [2,3], possess diverse expertise [6,7], and ultimately have different preferences on what they would like to contribute [1,3]. Furthermore, these users behave differently, depending on the tag or category (a sub-community interested on a specific topic) that they participate [1,8].

Studies of diversity in contributor's behavior are essential for understanding how each type of contributor collaborates for the functioning of Q&A systems. Site administrators can use these analyses to better manage

*Correspondence: adabriand@lsd.ufcg.edu.br
Federal University of Campina Grande, Av. Aprigio Veloso, 882 - Bloco CO, Bodocongo, Campina Grande, CEP 58429-900, Brazil

the site and to inform the design of task allocation mechanisms and personalized interfaces. Also, such studies are necessary for creating theories of how we behave in large collaborative groups online.

This paper contributes for the understanding of the collective production in Q&A sites by studying the typical contributors' behavior. In this work, we use historical data from the Super User Q&A site, a site with around of 66,000 contributors and 748,000, and devise a set of profiles that categorize contributors in two perspectives: one that considers all contributor's activity in the site and another that considers only their activity in tag communities.

The analysis of the first perspective reveals nine profiles, which can be summarized into four types: *no marked skill*, contributors with low to average quality contributions and activity levels; *unskilled answerer*, contributors with poorly evaluated answers; *experts*, contributors who are skilled in performing a kind of activity; and *activists*, highly active contributors. Based on these profiles, we found that the Super User community is composed mostly by contributors who fit into no marked skill and unskilled answerer profiles, and these contributors, together with activists, are responsible for the majority of the contributions in the site.

In a second analysis, we have evaluated the same user activities, but within sub-communities interested in specific topics. These communities are defined by the activity on questions marked with a specific tag, in this case, the most used operating systems tags. We found that the same nine profiles that describe Super User's usage patterns describe well the usage of such tag communities. We then compared community compositions and found that the evaluated communities have similar profile distributions, although each one has particularities that can be explained generally by the expected behavior of its users. This evidence suggests that administrative actions focusing in specific contexts such as tags can achieve better results.

Related work

Most studies on contributors' profiles in online communities have analyzed these profiles aiming to identify archetypical roles assumed by the users or to examine contributors which assume one profile known *a priori*, such as moderators or lurkers. In both cases, the overall goal is most often to help the interpretation of collective behavior or to inform the development of task allocation strategies or mechanisms to promote or inhibit certain behaviors.

Characterizing contributors' behavior

In this perspective, a large body of work examines the historical usage data, or conducts observational fieldwork, to identify salient contributor profiles (or roles) in online communities. Studies along these lines have been conducted in varied contexts, including newsgroups (e.g., Usenet [9-13]), Wikipedia [14-16] and other wikis (e.g., Cyclopath [17]), content-sharing communities [18,19], movie recommendation sites [4], and Q&A sites (e.g., Yahoo! Answers [1], Naver Knowledge-iN [3], and StackOverflow [2]). Although the profiles that best define the contributors are contingent of the community's context and the analysts' purpose, some regularities that are relevant to this work are discussed in the following. The profiles chosen to best describe a set of contributors to a community are dependent on the context of the community itself and on the purpose of the analysis. Nevertheless, there are regularities in the literature performing related analyses that are relevant to the present work, which are discussed in the following. In doing so, we refer to the contributors that provide high volumes of contributions as *highly active* and label the contributors that provide highly valued contributions as *expert* contributors.

From the perspective of how much contributors act, studies in several systems have generally found a majority of less active contributors collaborating with a small proportion of highly active contributors [2,3,13,15,17,19]. In the context of Q&A sites, highly active contributors have been found to represent 1% of the registered users but to account for 22% to 28% of all the answers in different sites [2,3].

The second part of the literature shows that contributions are often also skewed in their distribution over time. The level of activity of a contributor over time has been reported to typically display an initial burst in their level of activity followed by a marked drop in more than one system [15,17,20]. For Q&A systems, Mamykina et al. [2] and Nam et al. [3] found an intermittent behavior marked by inactive periods in the posting and answering behavior in different communities. Moreover, Nam et al. also observed that the time contributors are active is positively correlated with the quality of their answers. Regarding the dynamics of group behavior, Kittur et al. [15] found that the bulk of the contributions in Wikipedia and the social bookmarking site Delicious is gradually shifting from being provided by the highly active users to a product of the acts of less active contributors.

By distinguishing the types of contribution, it is possible to identify the groups of users that are highly active and/or experts only in certain types of contribution. Identifying the existing user groups in this perspective can, in turn, help manage the community and coordinate the contributors. Usenet has been thoroughly studied with varying methods, and the resulting literature consistently points to a description of newsgroup participants that include *answer persons*, *question persons*, and *trolls*, and *lurkers*

[9-12]. In the context of Wikipedia, distinct contributors of different profiles are chiefly responsible for certain types of contributions [16], and inexperienced and experienced users have been observed to contribute differently [14]. Welser et al. [16] also show that none of these profiles are formed by a large majority of experienced users. This observation suggests that the community does not depend heavily on experienced users for the service provided by editors in any of the profiles.

Specifically looking at Q&A sites, Nam et al. [3] suggested that, as in Usenet, contributors could be clearly separated as *askers* and *answerers* in Naver Knowledge-iN. Adamic et al. [1] also found these two profiles in Yahoo! Answers but observed as an additional prominent profile the *discussion person*, which is a user who is highly active in asking and answering questions.

Another trend in the contribution behavior examination is the study of social network structure created, for example, between askers and answerers. Kang et al. [21] examined the influence of heavy users' social networks on the quality of answers. Using data from Yahoo! Answers and Naver Knowledge-iN, the authors found that the number of askers helped by a user influences the quality of his/her answers and that heavy users have competing relationship among co-answerers.

Rodrigues et al. [8] studied sub-communities around the most frequent used question categories inside the Live QnA and the Yahoo! Answers systems to find that active users may establish strong social ties with categories. Finally, analyzing this last environment, Adamic et al. [1] characterized teh questions' categories, finding a large diversity of behavior when comparing different groups of categories (e.g., programming versus marriage) and that when dealing with factual expertise questions, more focused users tend to receive higher answering rates.

Examining or identifying experts and elite contributors

A different strand of research focuses on users from a specific profile. This research typically either examines such users closely or aims at deriving heuristics to predict which users will act according to the profile. The most popular profile considered in this approach is the highly active expert, sometimes named the 'elite' contributor. Understanding how such users behave and automatically identifying them can again improve task scheduling or recommendation and direct community efforts to fostering the participation of highly productive contributors.

Studies have shown that in some contexts, elite contributors have a consistent behavior from their start in the system [17,22]. Also, several algorithms have been devised to identify experts or authorities in certain themes in the community [23,24].

When looking at Q&A sites, most attention has been given to predicting which users are likely to be experts in certain themes. For example, Riahi et al. [25] and Hanrahan et al. [26] explore automatic means for identifying the most adequate experts for a question. Pal et al. showed that it is possible to predict which users will be highly active experts using data from their first weeks of activity [6,7]. Looking at the dynamics of expert behavior, Pal et al. [6] report that experts are either consistently active, initially inactive and later active, or the opposite.

Our contribution

This work advances the present knowledge about Q&A communities by using multivariate analysis techniques to characterize contributors' behavioral profiles. Each profile found represents a common co-occurrence of values in a set of metrics that describes the users' motivation (quantity) and ability (quality) for multiple types of contribution. By accounting for a more diverse set of metrics, this characterization enables us to investigate, for example, whether expert contributors are also highly active, and hence create large volumes of answers, questions, or comments. Such a richer picture can enable deeper understanding of different contributor skills, goals, and needs in Q&A sites (as also suggested before by Gazan [27]), and inform community management, experts' characterization, and task allocation.

Our previous work [28] has advanced this knowledge by applying multivariate techniques on Super User's community. In the present work, we both expand and improve this previous effort. First, we extend the generalization of our results of the first analysis in all Super User site by examining the contributor profiles in the context of several tag communities. Second, by revising the set of metrics, we redefined the metrics to achieve a simpler and more accurate model for the contributor profiles.

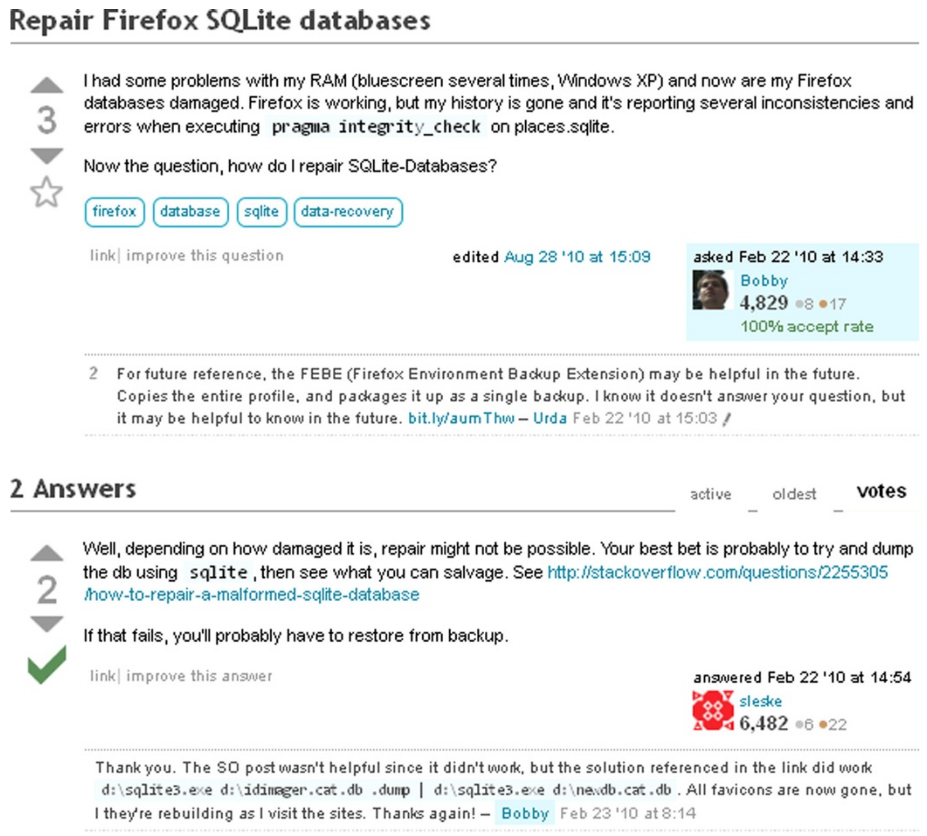
Site studied

Our analysis uses data from the second largest site of the Stack Exchange Q&A platform, Super User. This section describes how this platform works, the site studied, and the data used.

The Stack Exchange platform

Stack Exchange is a platform that allows the creation and hosting of Q&A sites. At the time of writing, the platform hosted 83 sites focused on topics varying from computer programming and theoretical computer science to culinary and photography.

Each of the sites operates independently and similarly to the Q&A model used by popular sites like Yahoo! Answers and Quora. Figure 1 shows the page of a question with their answers and comments in the Stack Exchange platform. In this Q&A model, the typical course of events for



Repair Firefox SQLite databases

I had some problems with my RAM (bluescreen several times, Windows XP) and now are my Firefox databases damaged. Firefox is working, but my history is gone and it's reporting several inconsistencies and errors when executing `pragma integrity_check` on places.sqlite.

Now the question, how do I repair SQLite-Databases?

2 For future reference, the FEBE (Firefox Environment Backup Extension) may be helpful in the future. Copies the entire profile, and packages it up as a single backup. I know it doesn't answer your question, but it may be helpful to know in the future. bit.ly/aumThw – Urda Feb 22 '10 at 15:03

2 Answers

Well, depending on how damaged it is, repair might not be possible. Your best bet is probably to try and dump the db using `sqlite`, then see what you can salvage. See http://stackoverflow.com/questions/2255305/how-to-repair-a-malformed-sqlite-database

If that fails, you'll probably have to restore from backup.

Thank you. The SO post wasn't helpful since it didn't work, but the solution referenced in the link did work `d:\sqlite3.exe d:\idimager.cat.db .dump | d:\sqlite3.exe d:\newdb.cat.db`. All favicons are now gone, but I they're rebuilding as I visit the sites. Thanks again! – Bobby Feb 23 '10 at 8:14

3
firefox
database
sqlite
data-recovery
link| improve this question
edited Aug 28 '10 at 15:09
asked Feb 22 '10 at 14:33
Bobby
4,829 ●8 ●17
100% accept rate
active
oldest
votes
2
link| improve this answer
answered Feb 22 '10 at 14:54
sleske
6,482 ●6 ●22

Figure 1 **Example of a question page in the Stack Exchange platform.**

a question posted in the site is (1) a user posts a question; (2) other users visualize the question and may vote its utility up or down, or favorite it; (3) one or more users post answers or comments associated with the question, which can themselves be visualized and voted up or down; and (4) the user who posted the question may at any point select one answer as the best answer.

As a result of this process, each question, answer, and comment has a voting balance (the number near the contributions) that is based on the positive and negative votes received, and questions also have a favorite count. Q&A sites usually use the voting balance of questions and answers to define how they will be ordered in the site.

Chosen site

The data of Super User site are described in Table 1. We choose Super User due to two reasons: it has the second largest community of Stack Exchange platform, and its history is more regular than the most popular site. The StackOverflow was the original instance from which Stack Exchange was generalized. As a result, it has a much longer and peculiar history compared to the remaining sites. The analysis of such changes on the platform's design is out of the scope of this work.

Table 1 Description of Super User as of July 2012

Site	Topic	Contribution	Posts	Creation
Super User	Computer power use	66,051	748K	2009-07

Data used

We use the historical data of Super User from its beginning until July 31st, 2012, as published by the Stack Exchange administrators. Stack Exchange periodically releases datasets describing the activity log of all registered users since the beginning of the site [29]. We processed this data and extracted the following two groups of metrics for each user.

The motivation metrics, in our context, are indicators of how motivated a contributor is in participating in the system. The ability metrics, in turn, aims to assess the quality of the content produced by a user. More details about these metrics are described as follows:

- Motivation metrics
  - *Number of questions* posted;
  - *Number of answers* posted;

  - *Number of comments* posted; and
  - *Activity duration*, defined as the number of days in which the user was active.

- Ability metrics

  - *Mean utility of questions (MUQuestions)*: gauges how the community perceived the quality of the user's questions. The quality of each question is measured as the sum of the number of favorites and its voting balance. A user's MUQuestions is the average utility of all questions posted by this user.
  - *Mean utility of answers (MUAnswers)*: measures the ability of the user in answering a question compared to that of competing answerers. For each answer that a user provided to a question and for which there are competitive answers, the utility of this user's answer is calculated by standardizing its voting balance compared to all other competing answers in that question (i.e., we calculate its z-score), or is set to zero if none of the answers have votes. A user's MUAnswers is the average utility of the answers posted by this user that had competing alternatives, or zero if none of the user's answers had competition. To better assess the quality of an answer, we add one positive vote in the calculation of voting balance if the answer was selected as the best.
  - *Mean utility of comments (MUComments)*: which evaluates how useful the community finds the comments a user makes in questions that were created by other users. It is calculated analogously to MUQuestions but considering the votes in comments posted by a user on other users' questions and answers.

Calculating the utility of answers by the z-score standardization avoids a bias of overestimating the answer quality caused by popular questions. For example, an answer given in a popular question could attract many votes, but this answer in comparison with other competing answers could be the one with less votes.

Methods

The problem of finding contributor profiles is analogous to identifying a set of groups of contributors with similar behavior. We use clustering analysis [30] to approach this problem. Clustering methods aim at finding a grouping solution that maximizes simultaneously in-group homogeneity and intergroup heterogeneity.

For this analysis, we use the set of motivation and ability metrics we defined and calculated these metrics, considering the complete activity of each user. Contributors that were not active before the last month in the dataset are excluded due to the small amount of information available about their behavior.

To define the space of similarity among users, we use the standardized values (z-scores) of the seven metrics considered to describe the contributor behavior. The similarity between the behaviors of two contributors is then the Euclidian distance between the contributors in the space defined by the standardized metrics.

Clustering algorithm

We employ a combination of hierarchical and nonhierarchical clustering algorithms to identify contributor profiles in our data. On the one hand, hierarchical clustering algorithms have the advantage of being independent of initial parameters (the number of clusters and their initial centers), and these algorithms allow the analyst to investigate a range of clustering solutions produced through iterative optimal cluster joining or splitting. Nonhierarchical clustering algorithms, on the other hand, optimize for global solution and provide solutions that are more robust to outliers than those of the hierarchical methods [31] but whose quality depends on an initial seed of cluster centers and presumes a known suitable number of clusters to be discovered.

Our analysis combines these two approaches by first using the Ward clustering algorithm [32] to explore a wide range of solutions with different numbers of clusters. The results of this exploration then inform a suitable number of clusters and their centers, which are in turn used to seed cluster centers in the k-means nonhierarchical algorithm [33]. We note that both the Ward algorithm and k-means are hard clustering techniques. Thus, each contributor is contained in exactly one cluster in our results.

Contributor profiles in the Super User

The first part of our analysis focuses on exploring salient contributor profiles in the Super User site.

Defining the number of clusters

The size of Super User's community (around 66,000 contributors) makes it impractical to run the Ward algorithm, because this method demands the computation of a similarity matrix $\mathbf{N}\times\mathbf{N}$. To address this issue, we execute the Ward algorithm on a random sample of 30,000 contributors (45.4%) to estimate the number of groups and their centers and use this information to execute a more scalable clustering algorithm on a second step.

Figure 2 displays the average within-cluster distances in a range of solutions that result from running the Ward algorithm in our data. This graph shows that the heterogeneity measure reaches a considerable stability from the 15-cluster solution onwards. Figure 3 also supports this

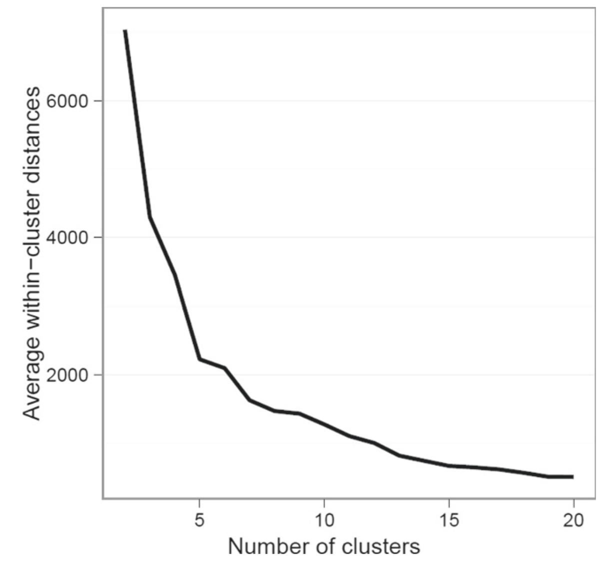


Figure 2 Analysis of the heterogeneity obtained in each cluster solution.

statement, showing that the percentage of heterogeneity reduction drops to 3% in the 16-cluster solution and that it continues to change around 0.3% to 10%.

Table 2 displays the centers from 15-cluster solution and a new center found in the 16-cluster solution. Examining these centers, we identify that the 15-cluster solution reveals a highly descriptive new cluster, while the new cluster from the 16-cluster solution is very similar to another already identified. In this work, a cluster is considered as highly descriptive if it has a center very distinct from the mean contributor behavior – a high distance (dissimilarity) to the metrics mean point (z-scores equal to zero).

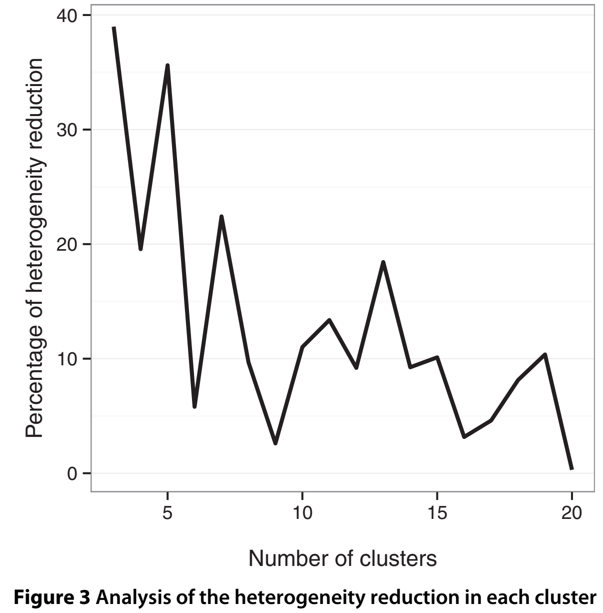


Figure 3 Analysis of the heterogeneity reduction in each cluster solution.

Choosing the 15-cluster solution and examining their centers, we identify the clusters that are very similar to each other. Therefore, in order to obtain a cluster center set with clearly distinct behaviors to seed the k-means algorithm in the next step, we manually grouped the cluster centers based on the similarity between them (Euclidian distance) and eliminate the centers of the less descriptive clusters from each group.

Table 2 also shows the six centers excluded from the 15-cluster solution. Among the candidate centers to be eliminated, we keep centers 1 and 2 because both have similar distance to the metrics mean point and each center has a peculiar variation of MUQuestions metric. Moreover, we also keep centers 11 and 12 due to the high variation of their metrics.

This elimination process results in a set with nine cluster centers that are the most distinct from each other. Yet, these cluster centers bear some similarity to the common behavior of groups of users well-known in the Q&A literature [1-3] (further discussions in the 'Results and discussion' section).

Finally, informed by this exploration, we use the algorithm k-means to identify nine clusters, and we seed this algorithm with the nine most relevant cluster centers identified in the 15-cluster solution of the hierarchical algorithm. The resulting clusters are similar to the most relevant in the hierarchical solution and are described in Figure 4. Appendix 1 also compares their centers using the unstandardized metric values.

Labeling the contributor profiles

Given the selected set of clusters, the next step in the analysis is to make sense of them in the context of Q&A sites. Our labeling of the nine identified profiles and their marked characteristics are as follows:

1. *Low-activity*: contributors with infrequent participation in the site and below-average motivation and skills.
2. *Occasional*: users that contribute moderately, and mainly questions, over an above-average activity time. Their questions tend to be considered useful.
3. *Unskilled answerer*: contributors of poorly evaluated answers, usually with low activity time. These users had not demonstrated any skill in providing answers.
4. *Expert answerer*: users whose numbers of postings are not pronounced but whose answers are consistently well evaluated.
5. *Expert questioner*: contributors whose questions the community recognizes as important and who are slightly more active than expert answerers.

Table 2 Standardized centers from 15-cluster solution and a new center found in the 16-cluster solution

Centers	Answers	Questions	Comments	Activity duration	MUAnswers	MUQuestions	MUComments	Distribution to the mean point
Center 1	-0.10	-0.08	-0.07	-0.15	0.14	*0.69*	-0.20	*0.76*
Center 2	-0.09	-0.20	-0.09	-0.18	0.17	*-0.47*	-0.20	*0.62*
Center 3[a]	-0.09	0.06	-0.06	-0.10	0.19	0.01	-0.1	0.31
Center 4	-0.07	-0.28	-0.09	-0.19	*-1.82*	-0.44	-0.20	*1.92*
Center 5[a]	0.01	0.25	0.01	0.12	*-1.14*	0.36	-0.02	1.23
Center 6[a]	-0.05	-0.29	-0.09	-0.17	*-0.87*	-0.45	-0.20	1.06
Center 7	0.03	-0.12	0.00	0.06	0.27	-0.04	*7.10*	*7.11*
Center 8[a]	0.12	-0.02	0.06	0.22	0.18	0.19	*1.16*	1.21
Center 9	0.06	0.02	0.03	0.11	0.12	*9.71*	0.77	*9.75*
Center 10[a]	-0.07	-0.05	-0.05	-0.10	0.14	*2.77*	-0.05	2.77
Center 11	*2.93*	*1.07*	*2.43*	*4.67*	0.27	0.72	0.82	*6.22*
Center 12	*23.40*	*5.35*	*24.78*	*22.13*	0.45	1.02	0.79	*41.01*
Center 13[a]	0.14	*2.52*	0.34	*1.07*	0.29	0.42	0.09	2.81
Center 15 (new in 15-cluster solution)	*1.03*	*13.61*	*2.25*	*4.90*	-0.03	0.67	0.14	*14.69*
Center 14	-0.06	-0.13	-0.07	-0.10	*2.13*	-0.11	-0.13	*2.15*
New center in 16-cluster solution[a]	-0.06	-0.28	-0.09	-0.17	*1.71*	-0.45	-0.20	1.8

Each center identified by Ward algorithm is listed in groups formed based on their similarity (Euclidian distance). [a]The center was excluded from the final center set (initial seed for k-means). The italicized values indicate a distinct metric of the respective center.

6. *Expert commenter*: users with little activity, but who produce valuable comments in the eyes of other users.
7. *Q-A activist*: contributors who are highly active in the site, chiefly creating questions, and whose answering skills are slightly below average.
8. *Answer activist*: users with a long activity time and high numbers of postings, specially answers. Moreover, with generally well-evaluated answers.
9. *Hyperactivist*: contributors with a profile similar to that of an answer activist, but who contributed a disproportionate number of answers and comments to the site and have the longest activity time among all profiles.

Results and discussion

The cluster analysis unveiled nine contributor profiles, which represent the most common combinations of activity level and skill among users from Super User site. Focusing on the less active profiles (all but activists and hyperactivists), we observe users that provide contributions of average quality, experts in one of each of the contribution types, and contributors that provide poorly evaluated answers. On the other hand, highly active users have less marked ability than their activity level. Figure 5 summarizes this view.

The evidence that experts and highly active contributors form disjoint groups contributes to understand where the expertise is located in the site, and it has implications for task allocation mechanisms. First, it suggests the need to examine whether present methods developed for identifying highly active experts (e.g., [6,7,25,26]) can accurately recognize the experts in our results. Second, it seems promising to use task allocation mechanisms to direct experts to primarily answer difficult questions when they contribute to the system. Finally, because of their low activity levels, task allocation mechanisms should consider suggesting a same question to multiple experts or to a combination of experts and highly active users to increase the chances of obtaining an answer in a reasonable time.

The observation that experts are typically less active contributors shows the need to further understand their motivations. On the one hand, if the site managers foster the participation of these users, it will likely have a positive impact on the service provided in the community. On the other hand, it seems necessary to understand to which degree these users are perceived as experts because they are very selective in the questions answered. It is not obvious that increasing the volume in their contribution will necessarily lead to mostly high quality contributions.

The salience of unskilled answerer profile draws the attention of site designers and managers. This calls for an investigation of the necessary means to reduce the potentially negative effect that such contributors may have in the site. Moreover, the absence of a distinct group of highly active unskilled answerer in our analysis suggests

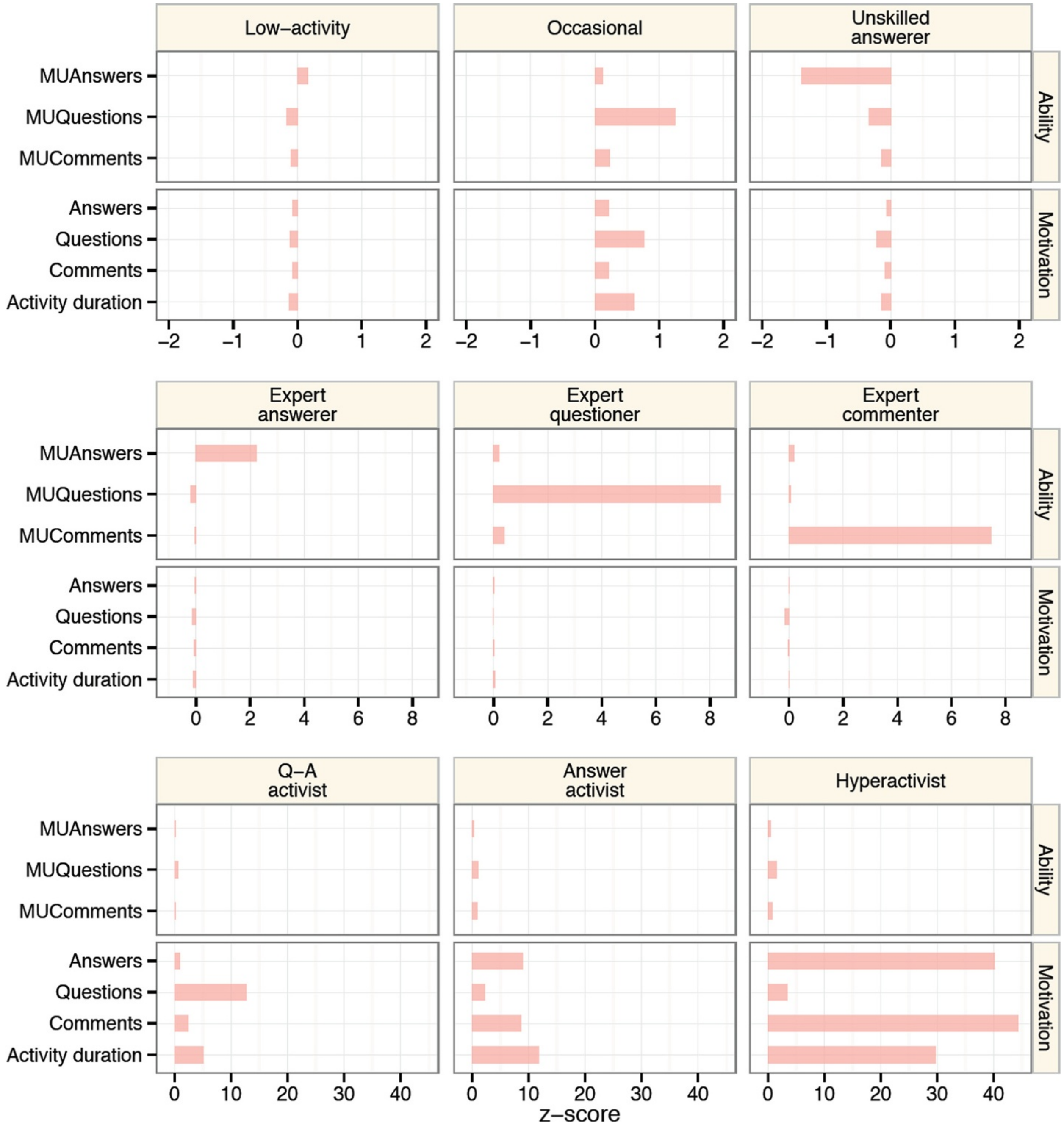


Figure 4 Cluster centers that describe the nine profiles in the analysis. Note that the horizontal axis is the *z*-score of a metric and that the scales in the three groups of graphs vary.

that the design of Stack Exchange inhibits the continued contribution of content perceived as poor. Examining which mechanisms have this effect and understanding how to better integrate them and improve the quality of their contributions can generalize good practices for Q&A sites.

The absence of highly active unskilled answerer can also be an effect of users migrating to another profile over time. These users can improve the quality of their answers or just stop to contribute due the negative feedback of the community. This hypothesis can be clarified by performing a longitudinal analysis of data to examine the contributors' dynamic behavior.

Among highly active contributor profiles, we notice that they distinguish themselves mostly by providing more questions than the average (Q-A activists), for answering

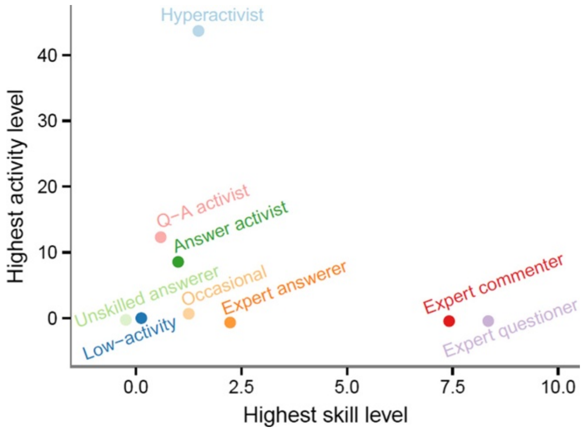


Figure 5 Scatter plot of the highest skill and activity levels defining the centroids of clusters found.

notably more than the average (answer activists), and for an exceedingly high contribution volume in general (hyperactivists). Identifying such users is relevant for allocating tasks in the community, as these are the contributors with the highest chance of providing a timely response to a request or suggestion. Also, a closer inspection at the hyperactivists highlights that these users are not only contributing content, but also concerned with the community functioning. An examination of the 2011 election for moderators in the Super User shows that one in six of the hyperactivists in that site nominated himself as a candidate. This number is orders of magnitude higher than the probability of any other profile volunteering in the election. Identifying hyperactivists can thus help community managers in finding users that can contribute in moderating the site.

Finally, our results are similar in some aspects to previous analysis of Q&A that looked into profiles according to contributors' activity levels [1,3]. Previous research also identified question- and answer-oriented profiles. However, our results complement this picture, considering the quality and quantity dimensions together, highlighting a more diverse set of profiles. Moreover, our analysis uncovers the unskilled answerer profile, which has not received much attention in the context of Q&A systems.

Site composition and profile productivity

In this section, we use the contributor profiles uncovered in the cluster analysis to examine the distribution of contributors among profiles in the Super User. Our goal is to investigate the prevalence of contributors in each profile and the role of user groups from each profile in producing knowledge in the site.

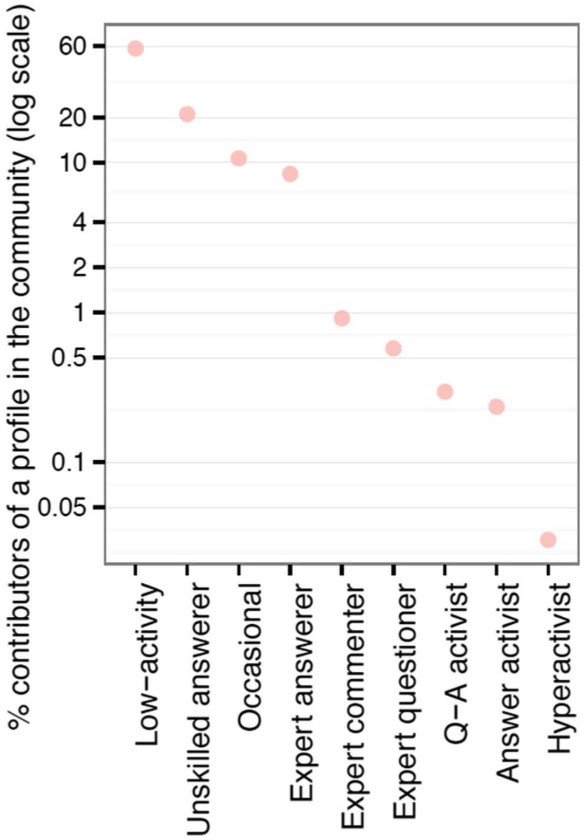


Figure 6 Distribution of the contributor profiles in the Super User.

Site composition

Figure 6 shows the distribution of the contributor profiles in the Super User. Slightly more than half of the contributors (57.8%) in each of the sites fit in the low-activity profile. The second and third most frequent profile are the unskilled answerers (21%) and occasional contributors (10.6%). Considering these three profiles, it is possible to note that the contributors of Super User are composed by 89.5% of low-activity, occasional, or unskilled users.

Among the remainder contributors, the most common profile is the expert answerer (8.4%), contributors with activity levels similar to low-activity users. Comparing experts and activists, we see that expert answerer occur more frequently than expert questioner, commenter, or activists of any type. Overall, it is not surprising that high-activity profiles are less common, these profiles demand more effort.

Profile productivity

Figure 7 contrasts the quantity and quality of contributions created by the groups of users with the different profiles in the Super User. For quantity, we consider how many questions, answers, and comments were produced

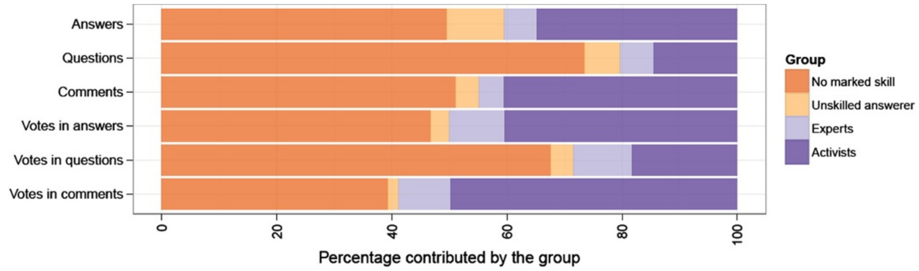


Figure 7 Parcel of the aggregate contribution for users in different profiles in the Super User. Profiles are grouped for readability, and the *no marked skill* label refers to low-activity and occasional contributors.

by contributors from each profile. For quality, we consider the number of positive votes received in answers, questions, and comments by the contributors. In this analysis, we refer to low-activity and occasional users together as contributors with no marked skill.

Both unskilled and expert users aggregately produce a small amount of contributions and receive a small proportion of the positive quality assessments in the site. Users with no marked skills and activists create the large majority of content. Surprisingly, the group of users with no marked skill produces more content than activists, and their contributions collectively receive most of votes in questions and answers. Furthermore, the fraction of created questions by no marked users is higher than their contribution in producing comments and answers. Conversely, highly active contributors aggregate more answers and comments than questions, and they receive most of votes in comments.

Our analysis reveals that the content in the Super User is mostly produced by two groups: a large base of contributors that act sporadically and have no marked skill, and a smaller more active nucleus of contributors. The first has a significant impact in generating questions, while the second has a remarkable importance in helping the community with quality comments. Experts and unskilled are of limited importance for the sheer number of contributions or votes received in contributions.

Regarding how answers are produced, our results are similar to those by Mamykina et al. [2]. They have found that highly active and low-activity contributors have similar importance in answer production in two other Q&A sites. Our results show how this pattern occurs also for questions and comments.

Finally, this characterization indicates that community managers should cater not only for activists, but also for low-activity and occasional users to keep the site productive. Along with highly active users, contributors with no marked skills are important for providing answers, questions, and comments. Also, our results point that the volume of content created by unskilled users is limited, but non-negligible. Mechanisms that help to improve the quality of their contributions may have a noticeable impact in the Super User.

Contributor profiles in tag communities

The Stack Exchange Q&A platform provides to each of its sites a tagging system to index content. Every time a question is created, the user marks it with one or more tags chosen from the set of already existing tags or if the user has enough qualification, he or she can create new tags. In this section, we use the same motivation and ability metrics (see 'Data used' section) concerning user's activity but focus the analysis on users and content from seven tags of the Super User site. Our goal is to get a deeper understanding of the site's composition and productivity by confronting them with the composition and productivity in the sub-parts of the site delimited by the tags. For this analysis and for the remainder of the paper, we refer to the set of users active in the questions labeled with a same tag as a *tag community*.

Tag communities contributor profiles

For our analysis on tag communities' composition, we selected the most prominent theme between the most used tags in Super User site: operating systems. Our intent is to limit the effect of the theme factor that has been shown to have influence on the community structure and production [1,8]. Moreover, we limit our sample to seven tags, as the next tags related to operating systems in the site have a number of contributors and posts orders of magnitude lesser than those selected. This selection also intend to limit the effect of the community size in our results. Table 3 summarizes the seven tag communities considered in this study.

Using the same clustering procedures described in 'Clustering algorithm' section, a hierarchical algorithm was executed over a random sample of contributors taken from each community with size equal to the smallest community (i.e., UNIX). The heterogeneity analysis of a range of cluster solutions again points to a solution with nine clusters. After this decision, the centroids of the nine clusters initially found were used to execute the k-means

Table 3 Operating systems tag communities considered in the case study

Tag	# Contributors	Posts
Windows 7	16,483	114,990
Linux	12,412	74,140
OS X	8,439	49,028
Ubuntu	7,795	36,014
Windows XP	6,979	39,467
Mac	5,573	26,878
UNIX	2,675	10,033

Posts are the sum of questions, answers, and comments.

algorithm over all contributors' activity data, aiming to adjust these clusters centers and to classify all users. The final centers obtained from this clustering analysis in the tag context (detailed in Appendix 2) are remarkably similar to the centers identified in our previous analysis in all sites (Figure 4).

Tag communities composition

After characterizing the nine prominent contributor profiles at the seven tag communities, we now examine how the communities are composed and how their composition varies. The tag community compositions are shown in Figure 8.

To assess variation in community composition, we test the independence between contributor profiles and the tag community they appear using a chi-square test. To meet the test's assumptions of a minimum number of observations per level of the user profile variable, the hyperactivist (H) profile was analyzed together with answer activist (AA) profile. The result of the test is a highly significant association between the usage profile occurrence and the tag community ($\chi^2(54) = 244.32$, $p < 0.001$), indicating that in spite of the similar overall pattern presented, the community compositions are significantly different.

To delve further into this analysis, we use the residuals information from the Chi-square, presented in Table 4. According to the residuals, the *Linux* and *Mac* tag communities' compositions are the most alike with the overall population composition, as no residual value shows significant differences. Following that, *OSX* and *Windows XP* have a composition also similar to the population average, but with a difference regarding one profile each.

From the most different communities, Windows 7 has a composition with significantly more low-activity users and less answer activists than the average for all tag communities. This particularity can be a reflex of a large community constituted by users that do not develop strong bonds with it. Furthermore, because this community discusses the system most often used by less advanced users, its questions may demand less skill to be answered in comparison with questions in other tags. As found in the residual data, users from this community tend to behave less as expert answerers and unskilled answerers.

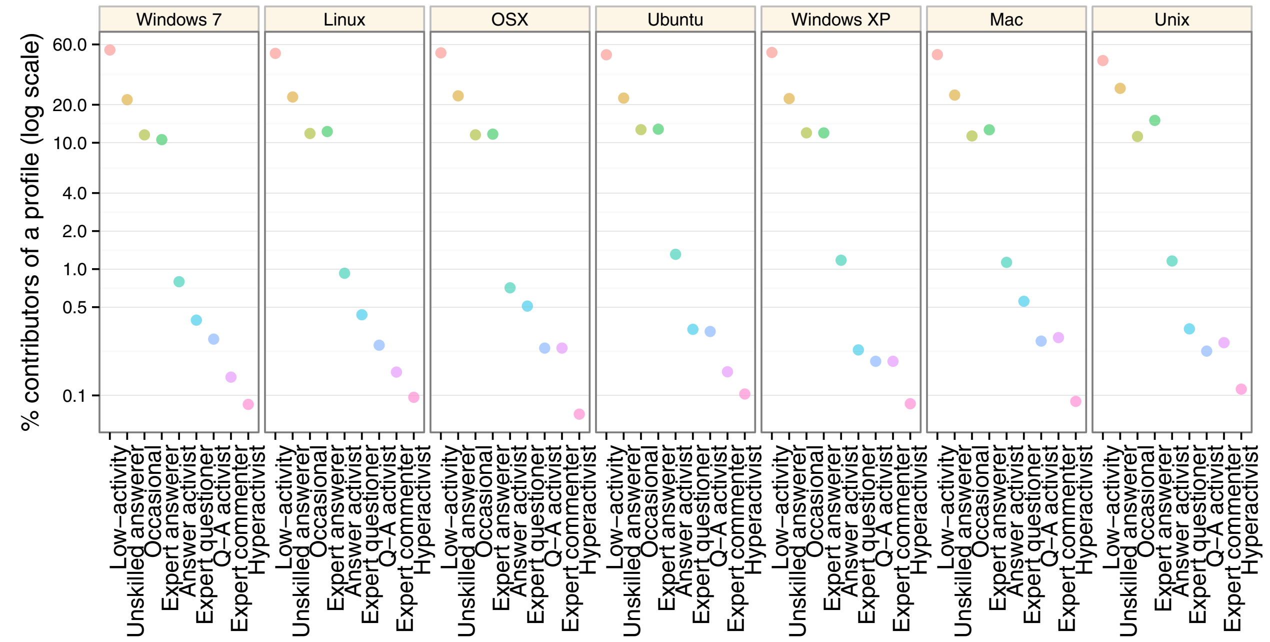


Figure 8 Tag community compositions (attention to the log scale in the vertical axis).

Table 4 List of chi-square residuals for tag community compositions dependency test

Profiles	Windows 7	Linux	OSX	Ubuntu	Windows XP	Mac	UNIX
Low activity	*4.934*	-0.744	-0.029	*-2.198*	0.383	-1.735	*-4.892*
Unskilled answerer	*-2.646*	0.28	1.16	-0.627	-0.94	1.468	*4.374*
Occasional	-0.848	0.232	-0.585	*2.39*	0.504	-0.978	-0.839
Expert answerer	*-4.967*	1.122	-0.604	*2.251*	0.101	1.612	*4.652*
Expert commenter	-1.285	-0.761	1.178	-0.585	0.079	1.833	0.962
Expert questioner	-0.2	0.54	1.521	-0.982	*-2.299*	1.784	-0.552
Q-A activist	0.52	-0.191	-0.388	1.081	-1.186	0.157	-0.348
Answer activist + hyperactivist	*-2.215*	-0.367	*-2.457*	*3.041*	1.656	1.184	1.076

Italicized values denote significant difference at $p < 0.005$.

In the UNIX tag community, we observe a significant negative residual for the low-activity profile. This can be explained by the common sense knowledge that UNIX users have a cultural tradition of being active in online forums and communities. With respect to the quality aspect, this tag has a significant positive residual for the expert answerer profile. This could be an effect of the theme discussed, which is one of the most specialized among the tags studied. Besides that, its questions demand more skills to be answered, which can explain the significant positive residual difference for the unskilled answerer profile.

Finally, the Ubuntu tag is the most popular Linux distribution with a large open source community, which may attract more active users. As a reflection of this, our results reveal that the answer activists and occasional profiles are significantly more frequent in this tag community. Concerning expert answerer and low-activity profiles, the same reasons considered for the UNIX tag community, a more specialized and dedicated group, can explain the remainder significant residual differences.

The fact that tag communities have different compositions suggests that strategies to promote and/or inhibit specific profiles can be more effective if applied in more specific contexts. In a case where a site administrator identifies a high amount of unskilled users, as observed in the UNIX tag community, he/she can focus on identifying (un)desired answering patterns to inform the development of mechanisms that alert and guide users to create better answers. Also, the high frequency of low-activity contributors observed in the Windows 7 community can motivate site.

Tag communities productivity

Given the profiles and the user's classification, it is possible to study how the group of profiles contributes to the overall content in the tag communities. In Figure 9, the contribution based on the number of answers, questions, comments, and votes is shown for each tag community. There is a remarkable similarity among all tag communities, and between them and the whole site (see Figure 7).

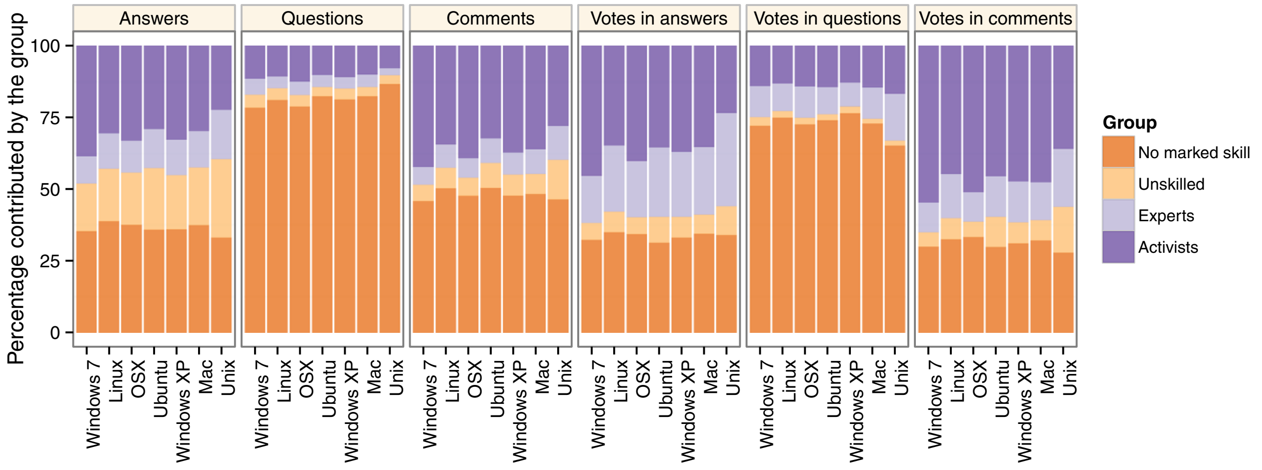


Figure 9 Contributions of tag community profile types.

Table 5 Unstandardized cluster centers of the nine profiles identified in the site analysis

Profiles	Answers	Questions	Comments	Activity duration	MUAnswers	MUQuestions	MUComments
Low-activity	1.06	1.16	1.87	2.46	0.00	0.64	0.04
Unskilled answerer	1.62	0.54	1.19	2.23	-0.85	0.27	0.03
Occasional	10.17	6.22	18.93	17.27	-0.02	3.88	0.20
Expert answerer	1.75	1.04	2.29	3.24	1.14	0.60	0.06
Expert commenter	2.98	0.90	4.14	5.28	0.01	1.15	3.58
Expert questioner	3.97	1.83	6.64	6.06	0.03	20.11	0.27
Q-A activist	33.83	75.32	147.81	104.78	-0.04	2.47	0.16
Answer activist	296.23	14.81	520.88	238.69	0.12	3.44	0.49
Hyperactivist	1,310.30	21.50	2,653.05	593.10	0.19	4.54	0.46

Besides the observed similarities, some differences can be noted when comparing the results for the whole site and the tag communities. On the one hand, the group with no marked skill seems more important to the site perspective when considering the amount of answers produced. On the other hand, unskilled answerers produce a higher portion of all answers in tag communities. Along with this, when considering the quality of answers, experts are more important to the tag communities (amount of received votes) than to the site as a whole.

Important differences could be observed too when comparing data between tag communities. In a broader view, only the largest (Windows 7) and the smallest (UNIX) communities are constantly different from the other five communities studied. The data shows that the Windows 7 community depends more heavily on the activists to produce answers and comments both in quantity and quality aspects; the UNIX tag community has the opposite tendency. Moreover, experts in Windows 7 are less important for the creation of answers and comments, while the UNIX community again shows the opposite trend. These observations can be also results of the typical contributors' behavior from each community, as discussed in 'Tag communities composition' section.

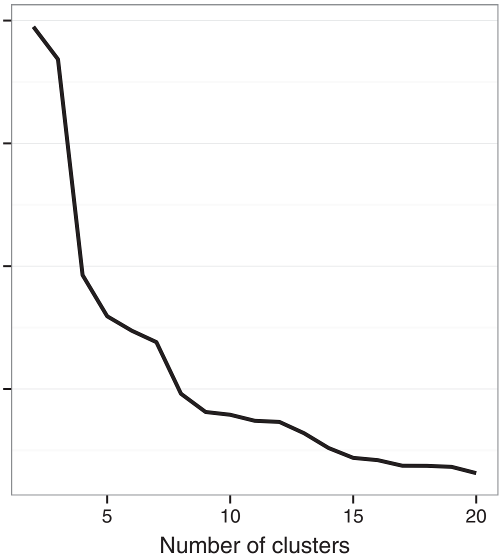


Figure 10 Analysis of the heterogeneity obtained in each cluster solution for tag analysis.

The differences found between the site and the tag communities reinforce the importance of analyzing the contributor behavior in specific contexts. For example, in the specific context of the tag communities we analyzed, we find that unskilled answerers produce a considerable amount of all answers. These answers are often of lower quality than the average, calling the attention of site administrators.

Conclusions

This work characterizes prominent contributor behaviors in the Super User site, considering how much and how well users contribute. Using the found profiles, we analyze the composition and productivity of groups formed by different types of contributors. Moreover, we contrast the whole site analysis with the analysis of seven of its tag communities.

The nine profiles uncovered enhance the overall understanding of how Q&A sites work. Such profiles can be grouped into four general types: no marked skill, unskilled answerers, experts, and highly active contributors. Knowing these profiles can support the development of management strategies in this and likely other sites. Particularly, finding that experts and highly active contributors are distinct groups of users is useful knowledge for the development of task allocation mechanisms and expert identification algorithms.

In the analysis of site composition and profile productivity, we observe that the Super User site is maintained mostly by a large base of contributors who acts

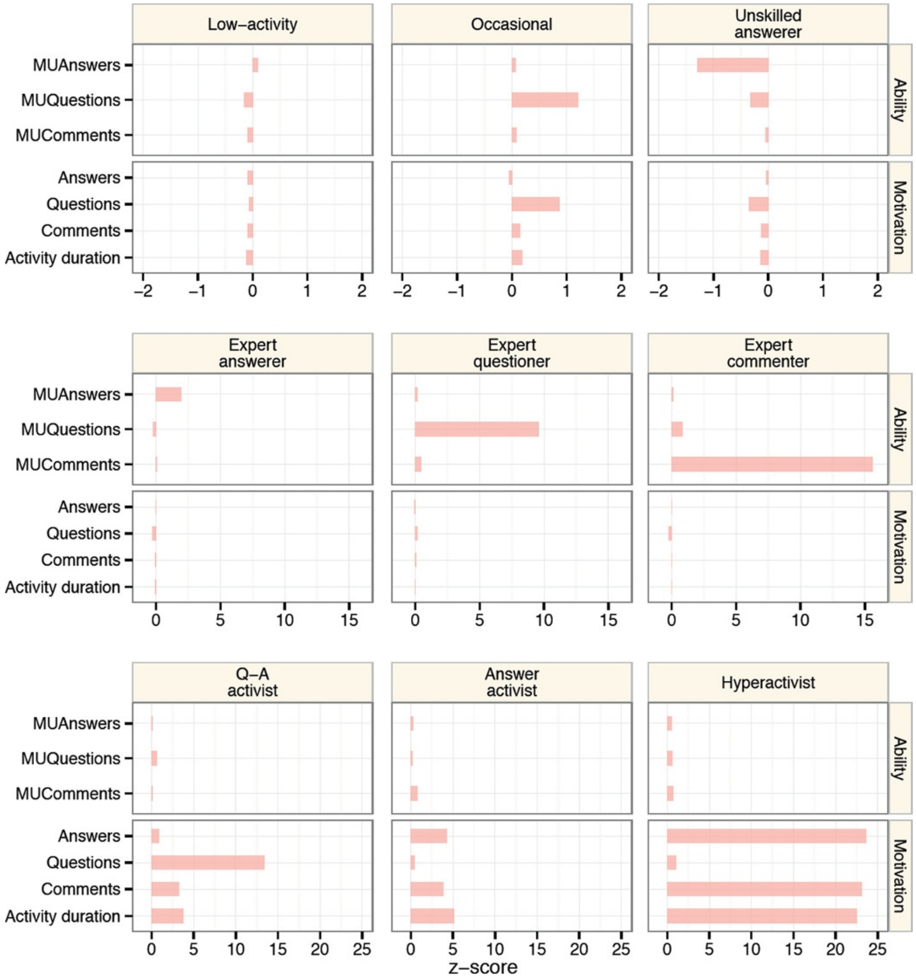


Figure 11 Usage profiles defined by cluster centroids in the tag analysis.

sporadically and have no marked skill, and a small group of active contributors. This portrait of the community highlights the importance not only of activists, but also of no-marked-skill users to generate content for the site.

Our analysis of productivity and composition also draws the attention to experts and unskilled answerers. The limited importance of experts in creating content can motivate site managers to foster their participation. The increased participation of these users may raise the overall

Table 6 Unstandardized cluster centers of the nine profiles identified in the tag communities

Profiles	Answers	Questions	Comments	Activity duration	MUAnswers	MUQuestions	MUComments
Low-activity	0.87	0.71	1.41	2.10	0.00	0.53	0.05
Unskilled answerer	1.44	0.14	0.77	1.80	-0.87	0.09	0.07
Occasional	1.25	2.57	5.24	5.53	-0.01	4.20	0.16
Expert answerer	1.70	0.27	1.58	2.53	1.17	0.28	0.15
Expert commenter	1.71	0.39	2.49	3.26	0.01	3.14	9.94
Expert questioner	1.01	1.71	3.54	3.44	0.05	26.62	0.37
Q-A activist	9.91	27.61	54.77	45.64	0.00	2.59	0.14
Answer activist	40.45	1.78	64.31	60.80	0.12	1.54	0.60
Hyperactivist	217.42	2.88	372.21	255.05	0.28	2.58	0.54

quality of contributions in the site. With respect to unskilled answerers, it seems promising to provide guidance in the creation of their answers. Since they are the second most common type of contributor, increasing the quality of their contributions can significantly improve the content produced in the site.

Regarding the analysis based on tag communities data, we overall find that the site exhibits remarkable self-similarity. The profiles needed to describe user activities in the tag communities are the same as those found for the site as a whole. This result improves our confidence on the generality of the found profiles as basis for analyses of Q&A sites' activity.

By evaluating the composition and productivity in all seven tag communities, we find an overall similar pattern, albeit with particularities. These particularities seem to be related to expected behavior, given the themes of the different tag communities. The differences between the tag communities suggest that the administration actions taken in specific contexts could increase the success rates of interventions.

Future work should focus on understanding the reasons behind the contributors' behavior. Our analysis points to the need to investigate the motivation of the users who fit in the expert and unskilled answerer profiles. Along with this investigation, a qualitative understanding of the remaining profiles could help create a clearer and richer picture of contributor behavior in Q&A systems.

Our work can also be improved by analyzing the relationship network created between users. This study can reveal if there exists trust relationships or other types of relationships between askers, answerers, and commenters.

Future work can also perform a longitudinal analysis to examine the contributors' dynamic behavior. This analysis can unveil the common transitions between the behavioral profiles and the contributors' evolutionary behavior, for example, how a newcomer becomes an expert or an activist.

Some aspects of our experiment pose threats to the validity of our results and should be improved in a future work. First, a new experiment is necessary to perform an external validation of the obtained clusters. It can verify how much our clustering method fails in classifying the contributor profiles. Second, new analyses are necessary to compare the efficiency of multiple clustering methods and of other metrics that could better describe the contributor behavior.

Finally, another direction in which our study could be expanded is the evaluation of a larger sample of Q&A sites and tag communities. This study would improve the generalization of the results for sites of different themes and in different platforms. Also, this experiment can correlate the community's composition with performance metrics, such as, the mean time for the community to solve a question. This investigation could reveal if a specific distribution of users in the contributor profiles reduces the site's performance.

Appendices

Appendix 1: unstandardized cluster centers from the first analysis in Super User

The unstandardized cluster centers of the nine profiles are shown in Table 5.

Appendix 2: results of the clustering analysis of the seven tag communities from Super User

Figures 10 and 11 and Table 6 show the details of the final centers obtained from the clustering analysis in the tag context.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AF focused to perform the cluster analysis over the site's activity data as whole, and NO extended this experiment to the tag communities. NA supervised and coordinated both studies. All authors read and approved the final manuscript.

Received: 21 December 2012 Accepted: 25 October 2013
Published: 23 January 2014

References

1. Adamic LA, Zhang J, Bakshy E, Ackerman MS (2008) Knowledge sharing and yahoo answers: everyone knows something. Paper presented in the 17th international conference on World Wide Web. Beijing, China, 21–25 April 2008
2. Mamykina L, Manoim B, Mittal M, Hripcsak G, Hartmann B (2011) Design lessons from the fastest Q&A site in the West. Paper presented in human factors, CHI '11. Vancouver, Canada, 7–12 May 2011
3. Nam K, Ackerman MS, Adamic L (2009) Questions in, knowledge in?: a study of Naver's question answering community. Paper presented at conference on human factors in computing systems, Boston, MA, USA, 4–9 April 2009
4. Fugelstad P, Dwyer P, Filson Moses J, Kim J, Mannino CA, Terveen L, Snyder M (2012) What makes users rate (share, tag, edit …)? Predicting patterns of participation in online communities. Paper presented at the conference on computer supported cooperative work, Seattle, WA, USA, 11–15 Feb 2012, pp 969–978
5. Tausczik YR, Pennebaker JW (2012) Participation in an online mathematics community: differentiating motivations to add. Paper presented at the conference on computer supported cooperative work, Seattle, WA, USA, 11–15 Feb 2012, pp 207–216
6. Pal A, Chang S, Konstan JA (2012) Evolution of experts in question answering communities. Paper presented at the sixth AAAI international conference on weblogs and social media, Trinity College in Dublin, Ireland, 4–7 June 2012, pp 274–281
7. Pal A, Farzan R, Konstan JA, Kraut RE (2011) Early detection of potential experts in question answering communities. Paper presented at the 19th international conference on user modeling, adaption, and personalization, Girona, Spain, 11–15 July 2011, pp 231–242
8. Rodrigues EM, Milic-Frayling N, Fortuna B (2008) Social tagging behaviour in community-driven question answering. In: IEEE (ed) Proceedings of the IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, vol. 1. IEEE, Washington, pp 112–119
9. Fisher D, Smith M, Welser HT (2006) You are who you talk to: detecting roles in usenet newsgroups. HICSS 3: 59b
10. Golder SA, Donath J (2004) Social roles in electronic communities. Internet Res 5: 1–25

11. Turner TC, Smith MA, Fisher D, Welser HT (2005) Picturing usenet: mapping computer-mediated collective action. J Comput-Mediated Commun 10(4): 0. doi:10.1111/j.1083-6101.2005.tb00270.x
12. Welser HT, Gleave E, Fisher D, Smith M (2007) Visualizing the signatures of social roles in online discussion groups. J Soc Struct 8(2): 1–32
13. Whittaker S, Terveen L, Hill W, Cherny L (1998) The dynamics of mass interaction. In: Lueg C, Fisher D (eds) Proceedings of the ACM conference on computer supported cooperative work. ACM, New York, pp 257–264
14. Bryant SL, Forte A, Bruckman A: Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. Paper presented at the international ACM SIGGROUP conference on supporting group work Sanibel Island, FL, USA, 6–9 Nov 2005, pp 168–176
15. Kittur A, Chi E, Pendleton BA, Suh B, Mytkowicz T (2007) Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. Algorithmica 1(2): 1–9
16. Welser HT, Cosley D, Kossinets G, Lin A, Dokshin F, Gay G, Smith M (2011) Finding social roles in Wikipedia. Paper presented at the the 2011 iConference, Seattle, WA, USA, 8-11 Feb, pp 122–129
17. Panciera K, Priedhorsky R, Erickson T, Terveen L (2010) Lurking? Cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. Paper presented at the SIGCHI conference on human factors in computing systems, Atlanta, GA, USA, 10–15 April 2010, pp 1917–1926
18. Andrade N, Santos-Neto E, Brasileiro F, Ripeanu M (2009) Resource demand and supply in BitTorrent content-sharing communities. Comput Netw 53(4): 515–527
19. Font F, Roma G, Herrera P, Serra X (2012) Characterization of the Freesound online community. Paper presented at the third international workshop on cognitive information processing, Baiona, Spain, 28 May 2012
20. Guo L, Tan E, Chen S, Zhang X, Zhao YE (2009) Analyzing patterns of user content generation in online social networks. Paper presented at the 15th ACM SIGKDD international conference on knowledge discovery and data mining, Paris, France, 28 June–1 July 2009, pp 369–378
21. Kang M, Kim B, Gloor P, Bock G-W (2011) Understanding the effect of social networks on user behaviors in community-driven knowledge services. J Am Soc Inf Sci Technol 62(6): 1066–1074
22. Panciera K, Halfaker A, Terveen L (2009) Wikipedians are born, not made: a study of power editors on Wikipedia. Paper presented at the ACM 2009 international conference on supporting group work, Sanibel Island, FL, USA, 10–13 May 2009, pp 51–60
23. Pal A, Counts S (2011) Identifying topical authorities in microblogs. Paper presented at the fourth ACM international conference on web search and data mining, Hong Kong, China, 9–12 Feb 2011, pp 45–54
24. Agarwal N, Liu H, Tang L, Yu PS (2008) Identifying the influential bloggers in a community. Paper presented at the international conference on web search and web data mining, Palo Alto, CA, USA, 11–12 Feb 2008, pp 207–218
25. Riahi F, Zolaktaf Z, Shafiei M, Milios E (2012) Finding expert users in community question answering. Paper presented at the 21st international conference companion on world wide web, Lyon, France, 16–20 April 2012, pp 791–798
26. Hanrahan BV, Convertino G, Nelson L (2012) Modeling problem difficulty and expertise in stackoverflow. Paper presented at ACM 2012 conference on computer supported cooperative work companion, Seattle, WA, USA, 11–15 Feb 2012, pp 91–94
27. Gazan R (2011) Social Q&A. J Am Soc Inf Sci Technol 62(12): 2301–2312
28. Furtado A, Andrade N (2011) Ativistas, passageiros, ocasionais e especialistas: perfis de usuário na construção de um site de Q&A In: Paper presented at VIII Simpósio Brasileiro de Sistemas Colaborativos. Paraty, RJ, Brazil, 5-7 October 2011
29. Saffron S (2012) Stack exchange data explorer 2.0. http://blog.stackoverflow.com/category/cc-wiki-dump/. Accessed September 3rd, 2012
30. Aldenderfer MS, Blashfield RK (1984) Cluster analysis. Sage University paper series on quantitative applications in the social sciences, vol. 07-044, Sage, Beverly Hills
31. Milligan GW (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45(3): 325–342
32. Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301): 236–244
33. Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stat 28(1): 100–108

doi:10.1186/1678-4804-20-5
Cite this article as: Furtado: **A case study of contributor behavior in Q&A site and tags: the importance of prominent profiles in community productivity.** *Journal of the Brazilian Computer Society* 2014 **20**:5.