Contributor Profiles, their Dynamics, and their Importance in Five Q&A Sites

Adabriand Furtado, Nazareno Andrade, Nigini Oliveira, Francisco Brasileiro
Systems and Computing Department
Universidade Federal de Campina Grande – Brazil
{adabriand, nigini}@lsd.ufcg.edu.br, {nazareno, fubica}@computacao.ufcg.edu.br

ABSTRACT
Q&A sites currently enable large numbers of contributors to collectively build valuable knowledge bases. Naturally, these sites are the product of contributors acting in different ways – creating questions, answers or comments and voting in these –, contributing in diverse amounts, and creating content of varying quality. This paper advances present knowledge about Q&A sites using a multifaceted view of contributors that accounts for diversity of behavior, motivation and expertise to characterize their profiles in five sites. This characterization resulted in the definition of ten behavioral profiles that group users according to the quality and quantity of their contributions. Using these profiles, we find that the five sites have remarkably similar distributions of contributor profiles. We also conduct a longitudinal study of contributor profiles in one of the sites, identifying common profile transitions, and finding that although users change profiles with some frequency, the site composition is mostly stable over time.

Author Keywords
Q&A sites; Empirical Methods, Quantitative; Datamining and machine learning; Studies of Wikipedia/Web

ACM Classification Keywords
H.5.3 [Information Interfaces]: Group and Organization Interfaces – Computer-supported cooperative work

INTRODUCTION

Question and Answer (Q&A) sites are currently useful for thousands of people each day. In their archetypical form, these sites enable users to post questions, answers, and comments, and to provide feedback on the quality of these posts to collectively identify good questions, adequate answers and valuable comments. 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 [1,16,18].

It is only natural that the contributors of Q&A sites are diverse and behave accordingly. For example, users have different motivations to contribute [18,8,25], spend varying amounts of time contributing [16,18], possess diverse expertise [21,19], and ultimately have different preferences on what they would like to contribute [1,18].

As in any organizational context, examining the diversity in contributor behavior in Q&A sites is central for interpreting how these sites work internally. This interpretation can, in turn, help managing a site, inform the design of task allocation mechanisms, guide the development of personalized interfaces, and ultimately lead participants to better understand Q&A sites.

This paper contributes to the understanding of how contributors collectively produce knowledge in Q&A sites by examining the typical behaviors of these contributors. For that, we use historical data from five sites of the Stack Exchange Q&A platform, and devise a set of profiles that describes contributor behavior both in long- and short-term perspectives. In the long-term analysis, we leverage these profiles to compare contributor behavior in the five sites, and identify a number of regularities both in the composition of the sites, as well as in the aggregate contribution of groups of users from each profile. For the short-term analysis, we perform a longitudinal study in the most active of the five sites and investigate how user behavior changes over time. This analysis again reveals a noted stability on the composition of the site, and exposes the profiles that are most stable, as well as those associated with users that end up abandoning the site.

Our study advances previous knowledge in two major aspects. First, it considers a multifaceted characterization of contributors. The profiles we encounter are multivariate descriptions of contributors that simultaneously account for the quantity and quality of their questions, answers, and comments. Second, we investigate how stable user behavior is over time and the distribution of user profiles in a site changes over time.

CONTRIBUTOR PROFILES IN Q&A CONTEXTS

In the context of online communities, the analysis of contributor profiles has been largely done in two perspectives: to identify archetypical roles assumed by 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

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.

development of task allocation strategies or mechanisms to promote or inhibit certain behaviors.

Characterizing contributors' behavior

In the first perspective, there is a large body of work examining historical usage data or conducting 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 [29,26,6,30,10]), Wikipedia [28,5,15] and other wikis (e.g. Cyclopath [23]), content-sharing communities [7,4], movie recommendation sites [8], and Q&A sites (e.g. Yahoo! Answers [1], Naver Knowledge-iN [18] and StackOverflow [16]). Although the profiles that best define contributors are contingent of the community's context and the analysts' purpose, some regularities that are relevant for this work are discussed in the following. In doing so, we refer to contributors that provide high volumes of contributions as *highly active*, and dub 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 [16,18,30,23,15,7]. In the context of Q&A sites, highly active contributors have been found to represent 1% of registered users but to account for 22-28% of all answers in different sites [16,18].

A 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 initially burst in their level of activity followed by a marked drop in more than one system [15,23,11]. For Q&A systems, Mamykina et al. [16] and Nam et al. [18] 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 contributions in Wikipedia and the social bookmarking site delicious are gradually shifting from being provided by the highly active users to a product of the acts of less active contributors.

Differentiating types of contribution, it is possible to identify 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 managing the community and coordinating contributors. Usenet has been thoroughly studied with varying methods, and the resulting literature consistently points to a description of newsgroup participants that includes *answer persons*, *question persons*, *trolls*, and *lurkers* [29,26,6,10]. In the context of Wikipedia, distinct contributors of different profiles are chiefly responsible for certain types of contributions [28], and inexperienced and experienced users have been observed to contribute differently [5]. Welser et al. [28] also show that none of these profiles is 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. [18] suggested that, as in Usenet, contributors could be clearly separated in *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.

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 [23,22]. Also, several algorithms have been devised to identify experts or authorities in certain themes in the community [20,2].

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. [24] and Hamrahan et al. [12] 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 [21,19]. Looking at the dynamics of expert behavior, Pal et al. [19] report that experts are either consistently active, initially inactive and later active, or the opposite.

Our contribution

This work advances present knowledge about contributor profiles in Q&A sites in two main directions. First, it analyzes usage data from multiple Q&A sites to characterize contributors' behavioral profiles using multivariate analysis techniques. Each profile found in this characterization describes a common co-occurrence of values in a set of metrics that describes the quantity and quality of a contributor's activity for multiple contribution types. By accounting for the multiple contribution types and their quantity and quality dimensions, 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 [9]), and inform community management, experts' characterization, and task allocation.

Second, our analysis contributes to expand present knowledge about how contributors' behavior changes over time. Our present understanding of Q&A sites does not describe how the quantity *and* quality of different types of contribution by a user change over time, or how the site structurally evolves with respect to its contributors' profiles. This analysis is necessary to identify prominent trends in contributor behavior, which again have implications for community management, experts' identification and task allocation mechanisms.

SITES STUDIED

Our analysis uses data from five sites of the Stack Exchange Q&A platform. This section describes how this platform works, the sites 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 of sites like Yahoo! Answers and Quora. The typical course of events for a question posted in the site is: (i) a user posts a question; (ii) other users visualize the question and may vote its utility up or down, or favorite it; (iii) one or more users post answers or comments associated with the question, which can themselves be visualized and voted up or down; and (iv) 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, based on the positive and negative votes received, and questions also have a favorite count.

Chosen sites

The five sites studied are described in Table 1. It is important to note that our site selection is not random and does not aim at providing a representative sample of all Q&A sites in Stack Exchange. Instead, we purposefully select a set of Q&A sites that are mature and popular, have the sizes of their user bases in the same order of magnitude, deal with reasonably related topics, and have existed for a similar timespan. Our intent is to, on the one hand, find profiles that are not specific of a single site, and on the other hand limit the effect of factors other than the user population on the contributor profiles we find.

Notably, we exclude from our sample the most popular among Stack Exchange sites, StackOverflow. The main reason for this decision is that this site was the original instance from which Stack Exchange was generalized. As a result, it has a much longer and peculiar history compared to the selected sites. Such differences render a comparison of its contributors to those in other sites out of the scope of this work.

Table 1: Description of the five sites studied as of August 2011.

Site	Topic	Contrib.	Posts	Creation
Super User	Computer power use	43,775	278K	2009-07
Server Fault	System administration and support	37,649	246K	2009-04
Ask Ubuntu	Ubuntu	9,904	92K	2010-09
Programmers	Software development	13,927	54K	2010-07
Mathematics	Mathematics	6,568	56K	2010-07

Data used

We use the historical data of the five sites studied from their beginning until August 31st, 2011, as published by the Stack Exchange administrators. Stack Exchange periodically releases datasets describing the activity log of all registered users since the beginning of each site[1]. We processed this data and extracted the following two groups of metrics for each user, considering a period of activity:

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

[1] http://blog.stackoverflow.com/category/cc-wiki-dump/

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.

- *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 MUAnswers but considering the voting balance of comments posted by a user on other users' questions and answers.

CONTRIBUTOR PROFILES IN THE FIVE SITES

The first part of our analysis focuses on exploring salient contributor profiles in the five sites analyzed.

Method

The problem of finding contributor profiles is analogous to identifying a set of groups of contributors with similar behavior. We use clustering analysis [3] 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 calculate these metrics considering the complete activity of each user (we later revisit this procedure with a different timeframe). 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 prevent a bias towards behaviors most present in sites with larger user bases, we use a random sample of same size from users in each community. The number of active contributors in the smallest community, Mathematics, limits the size of this random sample. The resulting number of contributors taken from all communities is 32,840.

To define the space of similarity among users, we use the standardized values (z-scores) of the seven metrics considered to describe contributor behavior. Because the scales of the metrics in the different communities may differ, the standardization is performed by site. 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 advantages of being independent of initial parameters and allowing 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 the global solution and provide solutions that are more robust to outliers than those of hierarchical methods [17], 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 hierarchical clustering algorithm [27] 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 [13]. We note that both the Ward algorithm and k-means are hard clustering techniques. Thus each contributor is present in exactly one cluster in our results.

Defining the number of clusters

Figure 1 displays the average within-cluster distances in a range of solutions that result from running the Ward algorithm in our data. It is notable that solutions with more than 11 clusters have little homogeneity gain compared to the 11-cluster solution. Examining the 10 and 11-cluster solutions closer, we identify that in the 11-cluster solution, two of the resulting clusters are very similar, with their distinction being a little variation on contributor behavior. On the other hand, a highly descriptive cluster present in the 11-cluster solution is absent in the 10-cluster solution.

Informed by this exploration, we opt to use the k-means algorithm to identify ten clusters, and to seed the algorithm with the centers of the ten most relevant clusters identified in the 11-cluster solution of the hierarchical algorithm. The resulting clusters are similar to the most relevant in the hierarchical solution, and are described in Figure 2. Table 4 in Appendix A also compares their centers using the unstandardized metric values.

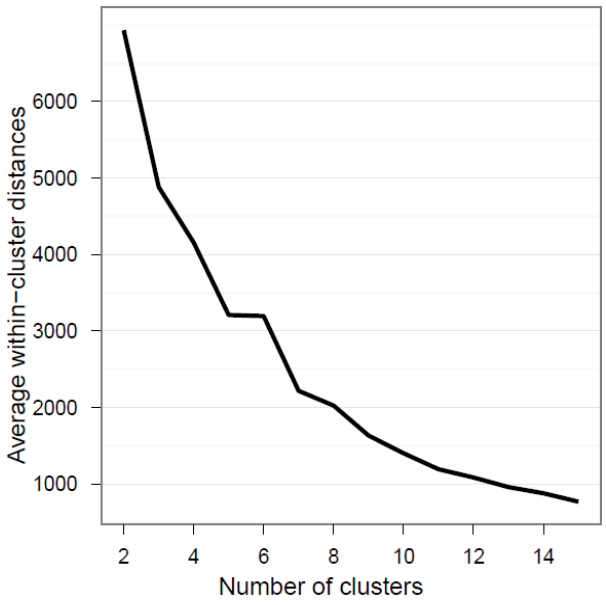

Figure 1. Analysis of the heterogeneity obtained in each cluster solution.

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 ten 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.

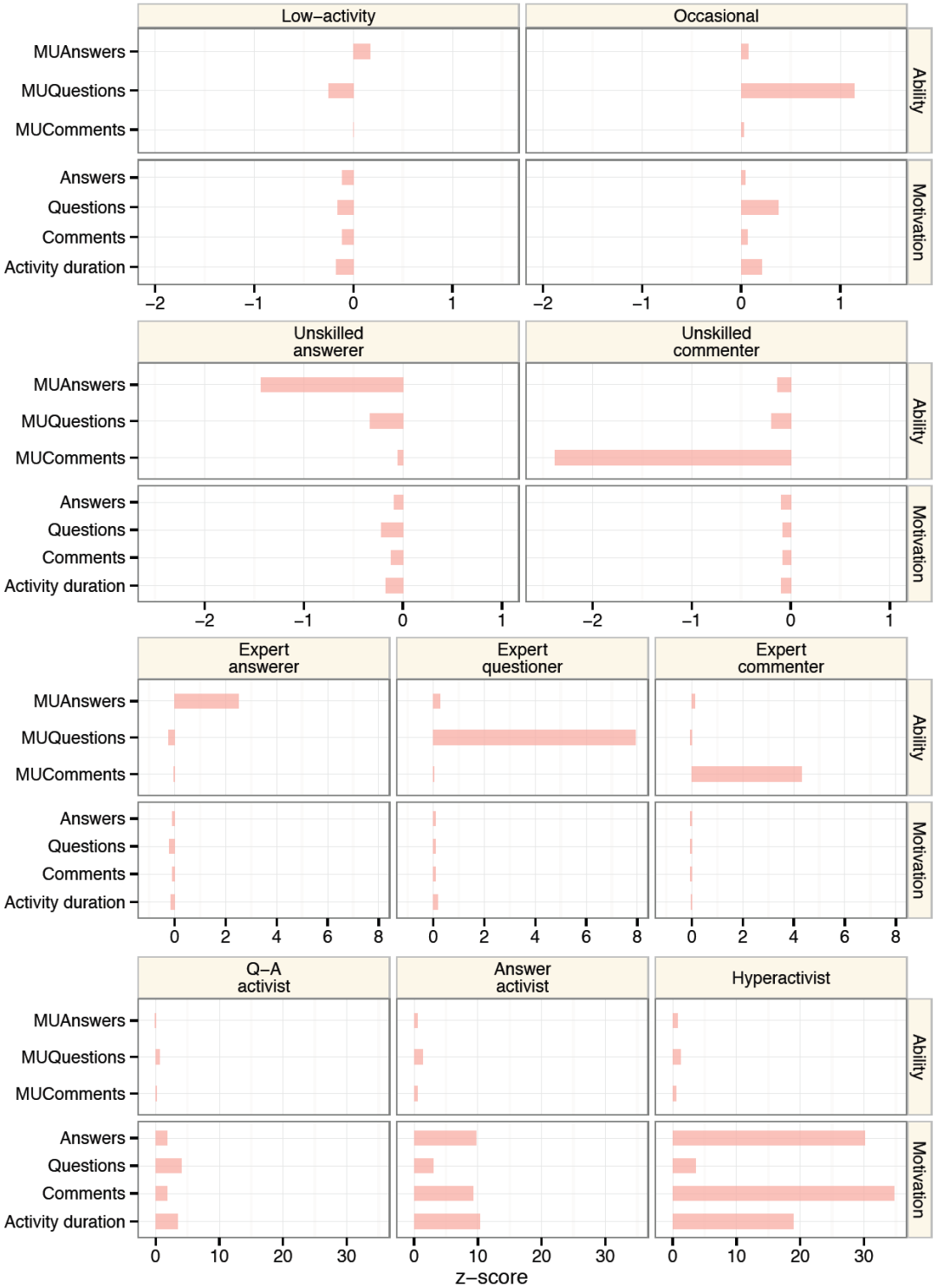

Figure 2. Cluster centers that describe the ten profiles in the analysis. Note that the horizontal axis is the z-score of a metric, and that the scales in the four groups of graphs vary.

3. *Unskilled answerer*: contributors of poorly evaluated answers, usually with low activity time. These users hadn't demonstrated any skill in providing answers.
4. *Unskilled commenter*: contributors that had comments rated substantially below those of other users commenting in the same questions.
5. *Expert answerer*: users whose numbers of postings are not pronounced, but whose answers are consistently well evaluated.
6. *Expert questioner*: contributors whose questions the community recognizes as important, and who are slightly more active than expert answerers.
7. *Expert commenter*: users with little activity, but who produce valuable comments in the eyes of other users.
8. *Q-A activist*: contributors who are highly active in the site, chiefly creating questions, and whose answering skills are slightly below average.
9. *Answer activist*: users with a long activity time and high numbers of postings, specially answers. Moreover, with generally well evaluated answers.
10. *Hyperactivist*: contributors with a profile similar to answer activists, but who contributed a disproportionate number of answers and comments to the site, and have the longest activity time among all profiles.

Discussion

The cluster analysis considering aspects of quantity and quality in users' contributions unveiled ten contributor profiles in the five Q&A sites studied. These profiles unveil the most common combination of activity level and skill among users. There are marked profiles defined by low to average, high and extreme activity levels. Among 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 or comments. Highly and extremely active users, in turn, are less marked by their ability than by their volume of contribution. Figure 3 summarizes this view. Notably, experts and highly active contributors are typically disjoint groups in our results.

The distinction between highly active contributors and experts is relevant for understanding where the expertise is located in the site, and has implications for task allocation mechanisms. First, it suggests the need to examine whether present methods developed for identifying highly active experts (e.g. [21,19,24,12]) 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.

From the management standpoint, observing that experts are typically less active contributors shows the need for further understanding the motivations of the highly skilled users. On the one hand, fostering the participation of these users 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 contributor profiles that our analysis identifies is also of interest to site designers and administrators. The absence of a distinct group of highly active unskilled users suggests that the design of Stack Exchange inhibits the continued contribution of content perceived as poor. Further examining which mechanisms have this effect can generalize good practices for Q&A sites. Nevertheless, in spite of the absence of highly active unskilled users, our results still point to marked profiles with consistently poorly evaluated contributions. This calls for an investigation of the necessary means to reduce the potentially negative effect such contributions may have in the sites.

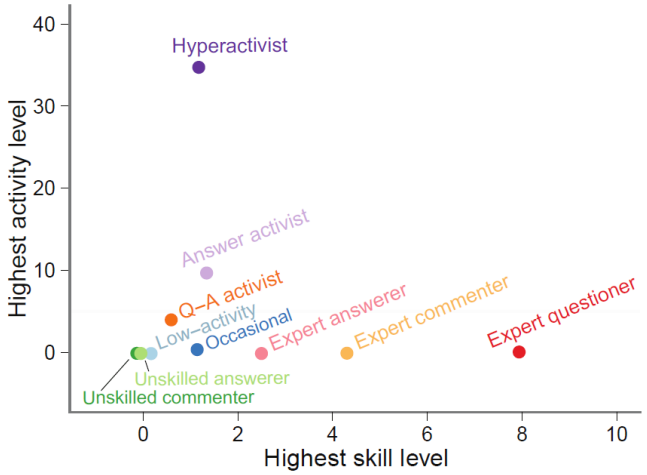

Figure 3 - Scatter plot of the highest skill and activity levels that define the centroids of the clusters found. Note that the points for unskilled answerer and unskilled commenter nearly coincide.

Focusing on the highly active users, we see profiles that distinguish themselves mostly by providing more questions than the average (Q-A activists), for answering 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 largest of the sites studied – 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 probability of any other profile volunteering in the election. Identifying hyperactivists can thus help community managers in finding users that can contribute to moderate the site.

In general, our results are in consonance with previous analyses of Q&A that looked into profiles according to contributors' activity levels [1,18]. Previous research also identified question- and answer-oriented profiles. Our results complement this picture considering quality and quantity dimensions together, highlighting a more diverse set of profiles. Moreover, our analysis uncovers the unskilled profiles, which have 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 five sites studied. Our goal is to investigate the prevalence of contributors in each profile, whether the five sites have similar compositions with respect to their contributors' profiles, and the role of user groups from each profile in producing knowledge in the different sites.

Finding all contributors' profiles

To analyze the contributor population in the five sites, it is necessary to find the profile of all users that were absent in the sample considered for the cluster analysis. For that, we train an artificial neural network (ANN) to classify users in profiles according to their motivation and ability metrics. The training data for the ANN is a sample of users whose profiles were identified in the cluster analysis. To facilitate the convergence of the ANN, this sample contains fewer cases from the more frequent profiles, and thus a higher proportion of the less represented ones. We find that the classifier has difficulties in distinguishing answer activists and hyperactivists, but considering these two profiles as one, the ANN achieves a 95% or higher detection precision for all profiles. We thus separate these classes in a second classification step, where a decision tree was trained only with data related to these two profiles in the training set. In this step the classifier achieves 95% or higher detection precision. Nevertheless, in the following analyses, we consider answer activists and hyperactivists together mostly because they seem to have the same impact on the sites.

Sites' composition

Figure 4 compares the distribution of contributors in profiles for the different sites. Generally, slightly more than half of the contributors in each of the sites fit in the low-activity profile. All sites also display similar proportions of occasional contributors (10-15%) and unskilled answerers (15-20%). Considering these three profiles, it is possible to note that the contributors of the five sites are typically composed by 80-90% of users with low to average activity and unskilled answerers.

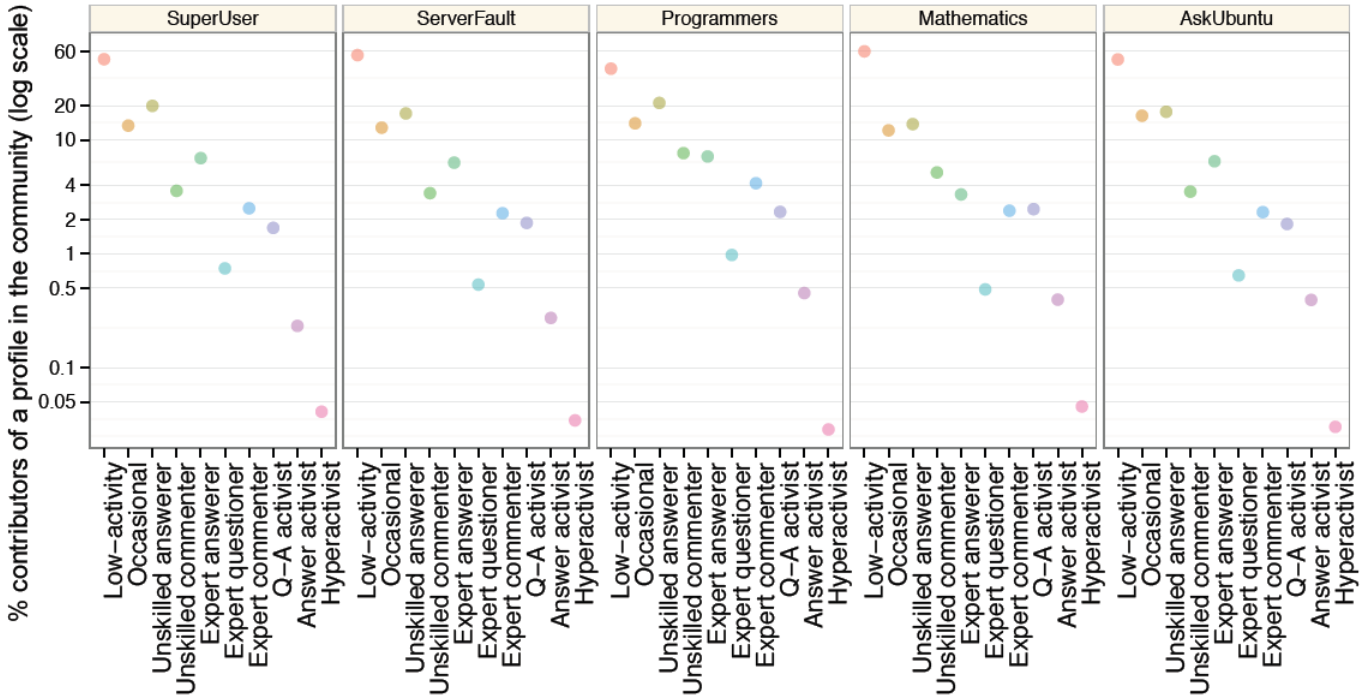

Figure 4. Distribution of the contributor profiles in the five communities.

Among the remainder contributors, the most common profile is the expert answerer. Comparing experts and activists, we see that expert answerers and commenters generally occur more frequently than expert questioners or activists of any type.

Overall, it is not surprising that expert and high-activity profiles are less common than those demanding less effort or skill. However, it is remarkable that, in spite of some peculiarities, the five sites exhibit the observed regularity in their composition.

Profile productivity

Figure 5 contrasts the quantity and quality of contributions from users of each profile in the five sites. For quantity, we consider how many questions, answers, and comments were produced 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.

In all sites, 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. Furthermore, highly active contributors are responsible for the largest part of answers and comments, and also receive most of the votes in these two types of contributions. Conversely, users with no marked skills contribute most of the questions, and receive more votes in their questions than the activists.

Overall, there is a visible regularity when comparing the results for the different sites. The noticeable exception is the Mathematics community, where activity and votes seem to be more concentrated on activists. This peculiarity may be related to the topic of the community, which may restrict more markedly the contribution of less dedicated users.

Considering the regularities, our analysis portraits the five sites as a combination of a larger base of contributors that act sporadically and collectively create most questions, and a smaller, more active nucleus of contributors that provides most of the answers to existing questions. These answers are complemented by contributors with no marked skill or unskilled, which still provide a sizeable proportion of all answers and receives a significant portion of votes in answers. Experts 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. [16]. They have found highly active and low-activity contributors to have similar importance in answer production. Our results show how this pattern occurs also for questions and comments, and is mostly regular across the five sites we consider.

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. Contributors with no marked skills are particularly important for providing questions. Also, our results point that the volume of content created by unskilled users is limited, but non-negligible. Mechanisms that help controlling the quality of contributions from these users may thus have a noticeable impact in these communities.

PROFILE DYNAMICS

Our analysis so far describes profiles and site compositions considering how the profile of a contributor is defined from the whole history of the contributor's activity. We now turn our attention towards a shorter-term characterization of profiles, investigating how the behavior of a contributor changes over time. Our goal is to examine the dynamics of contributor behavior and of sites' structural properties. For that, we use the largest of the five sites studied, Super User, and conduct a longitudinal version of our analysis in its historical data.

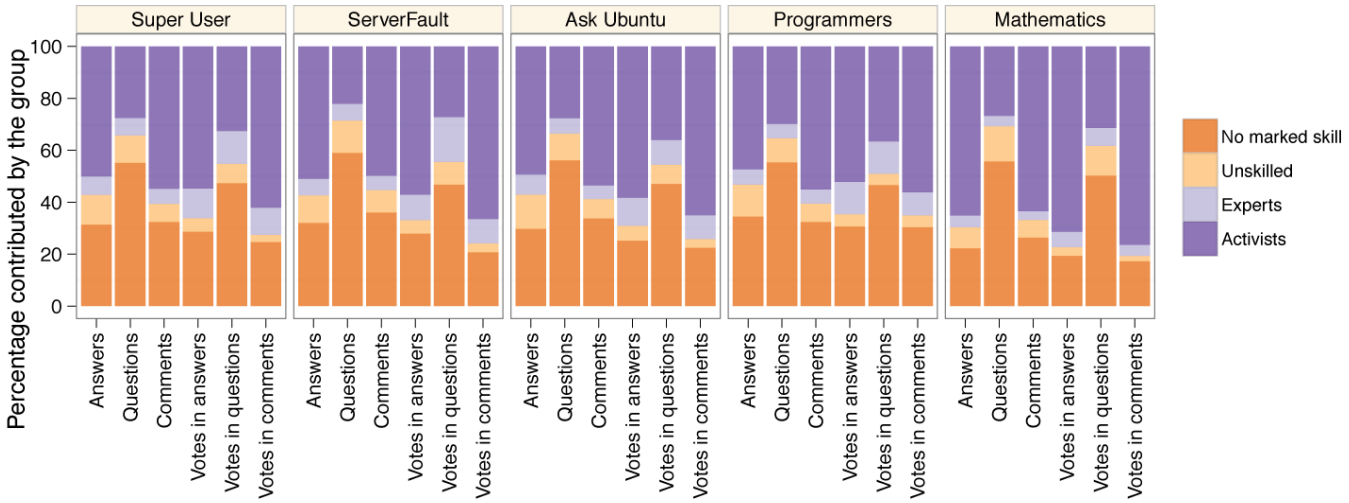

Figure 5. Parcel of the aggregate contribution for users in different profiles in the five communities. Profiles are grouped for readability, and the *no marked skill* label refers to low-activity and occasional contributors.

Method

The data for Super User encompasses a trace of 26 months of its activity. The mean (resp. median) time between the first and last seen contributions of users who start their activity in the first year of the trace is five (resp. eight) months. To avoid a bias towards identifying users departure from the system, even though they may contribute again in the following months, we exclude the last six months of the trace from the analysis. This period is only used to check if a user will still be active later than a given point in the trace. Using the remainder period in the trace, our analysis discretizes the data in ten two-month time windows. The size of the time window is chosen so that the activity of many contributors spans more than one unit of analysis.

A contributor is considered active in a time window if he or she contributes a question, an answer or a comment in the first month of that time window. Excluding contributors whose activity in the site start towards the end of a time window prevents us from underestimating the contributor's activity due to a too-short period of observation. Before performing any analysis, we also standardize contributors' motivation and ability metrics considering the activity of other contributors in the same time window.

After this preprocessing, we conduct a cluster analysis of active contributors using the same combination of hierarchical and nonhierarchical methods employed in the analysis of contributor behavior in the five sites.

Profile identification

The results of the cluster analysis for contributor activity considering 2-month time windows are remarkably similar to those of the analysis of the whole timespan of the site. The same ten profiles that characterized long-term contributor behavior are well suited to describe their short-term activity (Appendix B details the same results presented in the five sites analysis). It is worthwhile to note that identifying the same profiles in such different time windows and in different groups of contributors suggests the generalizability of the characterization results.

Structural dynamics of contributor profiles

The cluster analysis considering 2-month time windows informs us what was the distribution of contributor profiles over time in Super User. Figure 6 depicts how the proportion of contributors in each profile changes over time. The proportion of users in most profiles is stable over time. The noticeable exceptions to this stability are the proportion of low-activity users and expert answerers, and to a lesser extent, expert commenters, and unskilled commenters. Low-activity contributors increase in 10% in the second year of our data, while the experts mentioned consistently decrease over this period, with a sharper decrease for expert answerers. Unskilled commenters have a noticeable decrease in their proportion in the first year of the community.

Further examination shows that these trends are closely related to tendencies in the newcomers that the site attracts over time. Figure 7 shows the probability that a new contributor joining the community in different periods of its activity behaves according to the profiles whose probability change mostly in Figure 6. Over time, and more noticeably after the 6th time window, contributors that arrive in the community are to be more prone to behave as low-activity contributors, and less as the other profiles. Moreover, this trend starts one time window after the community begins to receive much higher numbers of new contributors than before (Figure 8).

The observation that the sustained high number of new contributors arriving in the community is correlated with a change in the profiles most often assumed by newcomers is of interest to community managers and merits further investigation. It is possible that Super User has started to attract a different type of public over time. However, it is also possible that the high number of arriving users has made it more difficult for new contributors to provide high-quality answers or comments, leading to the lower proportions of expert answerers and commenters in the community. Related to this result, it has been reported that Wikipedia has displayed a steady increase in the number of low-activity contributors over time [15].

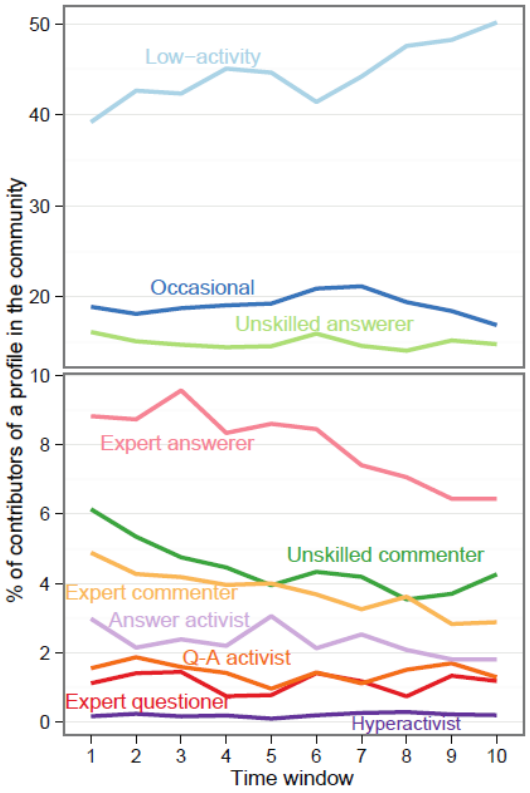

Figure 6: Composition of the population of active contributors in Super User over time. Note the scales are different in the two parts of the vertical axis.

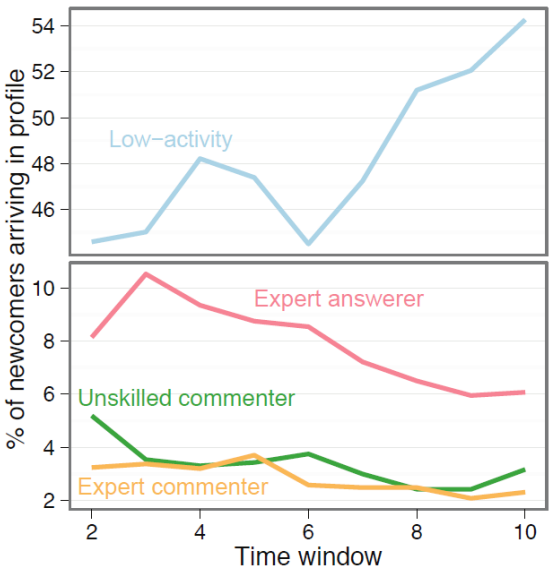

Figure 7: Percentage of newcomers in each window behaving according to the four profiles whose proportion change significantly over time in the community. Scales are different in the two parts of the plot.

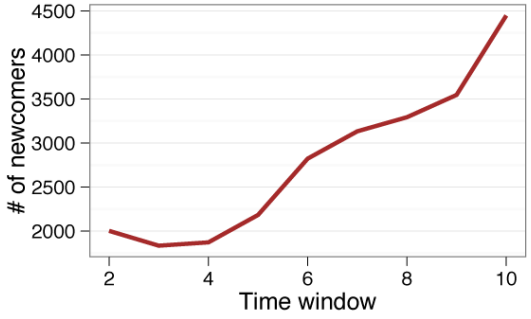

Figure 8: Number of newcomer contributors joining Super User over time.

Dynamics of individual contributor profiles

After examining the dynamics of site composition, we investigate the dynamics of each contributor's profile.

Our first experiment in this analysis is to measure the probability that a contributor transitions between any two profiles in two consecutive time windows. In this process, we also compute the probability that a contributor becomes inactive in the next time window, or abandons the site. A contributor is said to have abandoned the site if there is no contribution by that contributor neither in future time windows nor in the last six months of the dataset that are not analyzed as time windows. Inactive contributors are those that created no answers, questions or comments in a time window, but have not abandoned the system.

The transition probabilities between profiles, and from each profile to an inactive state or to abandoning the community are shown in Table 2. A contributor that is, at a moment in time, in any profile except activist or hyperactivist has a large tendency to reduce his or her activity in the following time window. Most of these contributors become inactive, low-activity or occasional contributors. Differently, activists and hyperactivists are the most stable profiles, and frequently maintain their behavioral profile in consecutive time windows. This observation is evidence that not only the proportion of highly active contributors is constant over time, but also that there is limited renovation among those behaving according to these profiles.

Also, this result reinforces previous analyses of the dynamics of activity in Q&A sites, which have consistently reported that some of the highly active contributors have a steady behavior, while most contributors have bursts of activity followed by inactive periods [16,18,19].

Further analyzing the highly active profiles, we see that hyperactivists and answer activists are closely related. There is a significant probability that hyperactivists turn into answer activists, and answer activists have the highest chance among all profiles of turning into hyperactivists. The dynamics of Q-A activists, on the other hand, are markedly different. There is little chance of a transition in either direction between this profile and answer activists or hyperactivists. Q-A activists are more likely to become occasional contributors, which indicates that activists that are not primarily concerned with answering and commenting have a more localized activity in time.

Considering again the overall picture, we see that besides having significant chance of becoming inactive for a complete time window, contributors are highly likely to remain in this state for consecutive periods. Among those returning to activity, most act moderately in the following time window, and low activity users, in turn, are likely to abandon the community.

Interestingly, the expert profiles show little stability. Moreover, expert answerers display a relatively high probability of abandoning the site. Similarly to the experts, contributors fitting in unskilled profiles are unlikely to remain acting in the same profile over time. Unskilled answerers show the highest chance of abandoning the site, and other unskilled contributors typically become low-activity or occasional users.

Taken together, our results point that expert profiles are not only more prevalent among less active users, but also that these experts tend to act during localized periods in the community's lifetime. Furthermore, expert answerers tend to act for a short period and abandon the community. This result indicates the need to further investigate how to increase the loyalty of such contributors. Also, the relatively short period of activity before abandoning the community implies that algorithms to identify experts should be able to locate these users early in their activity.

The observation that unskilled answerers have a high chance of abandoning the community suggests that providing poorly evaluated answers is demotivating to a significant proportion of contributors. On the one hand, this

Table 2: Probabilities of transitioning from profile X (row) to Y (column). As such, the sum of each row is 1. The CA column contains the probabilities of abandoning the community. Users that do not act in one time window are included as Inactive.

Profile	H	AA	QAA	EA	EQ	EC	UA	UC	O	LA	INA	CA
Hyperactivist (H)	.50	.37	.01	.01	.01	.00	.01	.03	.01	.01	.03	.00
Answer activist (AA)	.03	.47	.01	.06	.01	.03	.04	.04	.08	.18	.04	.01
Q-A activist (QAA)	.00	.04	.31	.02	.00	.02	.03	.01	.36	.14	.04	.02
Expert answerer (EA)	.00	.01	.00	.05	.00	.03	.05	.03	.07	.16	.28	.31
Expert questioner (EQ)	.00	.02	.01	.03	.01	.01	.05	.04	.11	.19	.28	.24
Expert commenter (EC)	.00	.01	.00	.07	.01	.06	.06	.05	.08	.20	.28	.18
Unskilled answerer (UA)	.00	.00	.00	.03	.00	.02	.06	.03	.06	.12	.23	.45
Unskilled commenter (UC)	.00	.01	.01	.05	.01	.04	.08	.05	.08	.20	.28	.19
Occasional (O)	.00	.01	.02	.03	.01	.02	.05	.02	.16	.20	.25	.24
Low-activity (LA)	.00	.01	.00	.03	.00	.02	.04	.02	.06	.16	.28	.38
Inactive (INA)	.00	.00	.00	.02	.00	.01	.04	.02	.06	.17	.66	-

Table 3: Results from a Pearson's chi-square test for association between being a newcomer or experienced contributor and the profile of the contributor. Hyperactivists and answer activists are considered together, to meet the assumptions of the test. Cells in bold denote the proportion found in the sample was significantly different from the proportion that would be expected if there was no association between the variables ($p < .05$).

Chi-square		H + AA	QAA	EA	EQ	EC	UA	UC	O	LA
Newcomer	Observed percentage	**1.37**	**0.90**	7.68	1.09	2.47	**18.29**	**3.02**	17.99	47.21
	Std residual	**-2.82**	**-2.01**	-0.44	0.01	-1.91	**2.26**	**-2.20**	-1.21	1.41
Experienced	Observed percentage	**4.17**	**2.50**	8.63	1.08	**4.90**	**10.88**	**6.13**	**22.01**	**39.71**
	Std residual	**9.05**	**6.45**	1.41	-0.03	**6.12**	**-7.26**	**7.05**	**3.87**	**-4.53**

phenomenon may limit the amount of content produced that does not meet the community standards. However, given the high proportion of unskilled answerers often present in the community (Figure 6), our results also indicate that Super User could benefit from improved mechanisms to educate its answerers. Preventing poorly evaluated answers may increase the retention of users willing to contribute to the site.

Newcomer vs. experienced contributor profiles

Our final experiment examines how the probability that a contributor behaves according to a given profile varies if the contributor is a newcomer or an experienced user. For the newcomer behavior, we examine the first profile assumed by all contributors. For the experienced contributor behavior, we use the fifth active profile of contributors that are active in at least five time windows. To test if there is an association between whether the contributor is a newcomer or not and the likelihood this contributor will behave according to a profile, we perform a Pearson's chi-square test. Table 3 shows the results of the comparison. Note that due to their small absolute number, hyperactivists were combined with answer activists to meet the assumptions of the chi-square test.

There is a significant association between being experienced or newcomer and the profile contributors display ($X^2(8)=330.0$, $p <$ 1e-10). Experienced users are more likely to be on more active profiles compared to newcomers. Their chance of behaving as answer activists/hyperactivists and Q-A activists are significantly higher, and they are more likely to behave as occasional and less likely to behave as low-activity contributors.

Considering the expert profiles, it is interesting that experienced contributors only have a significantly higher chance than newcomers to behave as expert commenters. This does not happen for expert answerers or questioners. Newcomers are, however, significantly more likely to act as unskilled answerers.

Surprisingly, experienced contributors are significantly more likely to behave as unskilled commenters than newcomers. This seems to be a side effect of their higher likelihood to comment on posts. During time windows in which these users concentrate their activity on commenting, they have a higher chance of adding accessory comments in discussions where other users add valuable comments than the newcomers. This explanation is reinforced by the fact that experienced users are also more likely than newcomers to act as expert commenters.

These results bear some similarity to observations by Welser et al. [28] on Wikipedia. However, while Welser et al. found that Wikipedia does not seem to depend on

experienced users for any role, we see that in Super User some profiles are more tightly related with experience.

Finally, the fact that newcomers are both as likely to act as expert answerers as experienced users, and more likely to act as unskilled answerers than the latter is also of interest. This result suggests that providing further guidance for newcomers in their first answers may increase retention in the community and lower the number of poorly evaluated content created.

DISCUSSION AND FUTURE WORK

This work provides a characterization of contributor behavior in five Q&A sites based on how much and how well users contribute different types of content over time. Building on this characterization, we analyze the composition and productivity of the sites, and examine the dynamics of contributor behavior over time in the largest of the five sites.

The ten profiles found in the characterization enrich our understanding of how Q&A communities work. Realizing that users that consistently provide high-quality contributions are not the highly active users is of particular interest. The knowledge that activists and experts are complementary profiles is useful for the development of task allocation mechanisms, expert identification and community management.

Regarding site composition, our study points to a high similarity in the occurrence of contributor profiles in the five sites. Moreover, the productivity of groups of users from the different profiles is also remarkably similar among sites. Although it is not possible to generalize these results for arbitrary Q&A sites, the regularities increase the chance that our findings are not specific of a site or context.

The analysis of composition and productivity also shows that unskilled answerers are a common occurrence in the communities. Adequately providing guidance and increasing the quality their contributions can thus have a significant effect on these sites. On a similar note, our results show that users with no marked ability produce a sizeable fraction of the content and receive a large proportion of the positive evaluations in the sites. This observation indicates that catering for such users has also a potential for the communities we study.

In the dynamic perspective, our work describe Super User, the largest among the sites studied, as a combination of a stable and highly active nucleus of contributors, a more volatile contingent of experts and unskilled users, and a large mass of low-activity contributors. This mass of low-activity contributors, in turn, includes at any time large proportions of the user that in other periods had a more marked activity. This result supports and complements multiple previous studies that report that the activity of contributors in Q&A sites happens in bursts.

Observing composition over time, we identify a trend in the increase of low-activity users combined with a decrease in expert answerers. Our analysis indicates that this tendency is linked to a change in the behavior of newcomer contributors, what merits further investigation.

Looking at the dynamics of behavior for individual contributors, we see that both unskilled answerers and the various expert profiles tend to abandon the community soon. This result reinforces the impression that there is an opportunity in identifying how to increase the retention of contributors likely to behave as experts, and how to better integrate the unskilled contributors.

Finally, our comparison of profile occurrence among newcomer and experienced contributors reveals that the community seems to depend on experienced users for behaving as activists. Interestingly, however, experienced users do not have a higher chance of acting as experts. This suggests that experts behave as such from their start in the system, but activists are likely to develop over time.

Future work should focus on deepening several aspects of our results. Investigating the motivations of users that fit in the expert profiles is of particular importance, but a qualitative understanding of all profiles unveiled in this analysis could help creating a clearer and richer picture of contributor behavior in Q&A systems.

A second line in which our results need further experimentation is on the evaluation of its generalizability. It is not clear if and how our results hold for Q&A sites of different themes and in different platforms. Finally, it may be fruitful to examine the dynamics of behavior in the different classes of contributors which we identify with a finer granularity. A longitudinal study of individuals in a contributor profile could also investigate the effect of system events on their productivity, and thus inform system designers and operators of risks for site productivity.

REFERENCES

1. Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. Knowledge sharing and yahoo answers: everyone knows something. *Proceeding of the 17th international conference on World Wide Beijing, C* (2008), 665-674.
2. Agarwal, N., Liu, H., Tang, L., and Yu, P. S. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, ACM (2008), 207–218.
3. Aldenderfer, M. S., and Blashfield, R. K. *Cluster Analysis*, vol. 07-044 of *Sage University Paper Series on Quantitative Applications in the Social Sciences*. Sage, 1984.
4. Andrade, N., Santos-Neto, E., Brasileiro, F., and Ripeanu, M. Resource demand and supply in BitTorrent content-sharing communities. *Comput. Netw. 53*, 4 (Mar. 2009), 515–527.

5. Bryant, S. L., Forte, A., and Bruckman, A. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Human Factors*, vol. 6 of *Net communities*, ACM (2005), 1–10.
6. Fisher, D., Smith, M., and Welser, H. T. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences HICSS06 00*, C (2006), 59b–59b.
7. Font, F., Roma, G., Herrera, P., and Serra, X. Characterization of the Freesound Online Community. In *Third International Workshop on Cognitive Information Processing* (Baiona, Spain, 2012).
8. Fugelstad, P., Dwyer, P., Filson Moses, J., Kim, J., Mannino, C. A., Terveen, L., and Snyder, M. What Makes Users Rate (Share , Tag , Edit ...)? Predicting Patterns of Participation in Online Communities. In *Communities*, CSCW '12, ACM(2012), 969–978.
9. Gazan, R. Social Q&A. *Journal of the American Society for Information Science and Technology 62*, 12 (2011), 2301–2312.
10. Golder, S. A., and Donath, J. Social roles in electronic communities. *Internet Research 5* (2004), 1–25.
11. Guo, L., Tan, E., Chen, S., Zhang, X., and Zhao, Y. E. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, ACM (2009), 369–378.
12. Hanrahan, B. V., Convertino, G., and Nelson, L. Modeling problem difficulty and expertise in stackoverflow. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, CSCW '12, ACM (2012), 91–94.
13. Hartigan, J. A., and Wong, M. A. A K-Means Clustering Algorithm. *Applied Statistics 28*, 1 (1979), 100–108.
14. Kang, M., and Kim, B. Understanding the Effect of Social Networks on User Behaviors in Community-Driven Knowledge Services. *Journal of the American Society for Information Science and Technology 62*, 6 (2011), 1066–1074.
15. Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Algorithmica 1*, 2 (2007), 1–9.
16. Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., and Hartmann, B. Design Lessons from the Fastest Q & A Site in the West. In *Human Factors*, CHI '11, ACM Press (2011), 2857–2866.
17. Milligan, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika 45*, 3 (1980), 325–342.
18. Nam, K., Ackerman, M. S., and Adamic, L. Questions in, knowledge in?: a study of naver's question answering community. *CHI* (2009), 779–788.
19. Pal, A., Chang, S., and Konstan, J. A. Evolution of Experts in Question Answering Communities. In *6th AAAI International Conference on Weblogs and Social Media*, ICWSM '12, The AAAI Press (2012), 274–281.
20. Pal, A., and Counts, S. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, ACM (2011), 45–54.
21. Pal, A., Farzan, R., Konstan, J. A., and Kraut, R. E. Early Detection of Potential Experts in Question Answering Communities. *Human-Computer Interaction* (2011), 231–242.
22. Panciera, K., Halfaker, A., and Terveen, L. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, ACM (2009), 51–60.
23. Panciera, K., Priedhorsky, R., Erickson, T., and Terveen, L. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. In *Human Factors*, CHI '10, ACM (2010), 1917–1926.
24. Riahi, F., Zolaktaf, Z., Shafiei, M., and Milios, E. Finding expert users in community question answering. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, ACM (2012), 791–798.
25. Tausczik, Y. R., and Pennebaker, J. W. Participation in an online mathematics community: differentiating motivations to add. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, ACM (2012), 207–216.
26. Turner, T. C., Smith, M. A., Fisher, D., and Welser, H. T. Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of ComputerMediated Communication 10*, 4 (2005), 0.
27. Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*, 301 (1963), 236–244.
28. Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. Finding social roles in Wikipedia. In *Methods*, vol. 48 of *iConference* '11, ACM (2011), 122–129.
29. Welser, H. T., Gleave, E., Fisher, D., and Smith, M. Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure 8,* 2 (2007), 1–32.
30. Whittaker, S., Terveen, L., Hill, W., and Cherny, L. The dynamics of mass interaction. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, CSCW '98, ACM (1998), 257–264.

APPENDIX A: UNSTANDARDIZED CLUSTER CENTERS FOR THE FIVE SITES STUDIED

Table 4: Unstandardized cluster centers for the profiles identified in the five sites studied.

Profile	Answers	Questions	Comments	Activity Duration	MUAnswers	MUQuestions	MUComments
Low-activity	1.14	1.06	1.47	2.73	0.00	0.87	0.01
Occasional	5.77	3.72	13.84	9.96	-0.05	7.46	0.02
Unskilled answerer	1.64	0.76	0.78	2.75	-0.84	0.47	-0.01
Unskilled commenter	1.62	1.47	3.50	4.23	-0.16	1.13	-0.70
Expert answerer	1.46	0.81	1.79	3.34	1.22	0.94	0.00
Expert questioner	6.94	2.26	15.23	9.16	0.05	39.73	0.01
Expert commenter	2.53	1.49	4.65	5.31	-0.03	1.78	1.29
Q-A activist	58.06	21.68	140.48	70.67	-0.11	4.86	0.04
Answer activist	299.49	16.91	668.80	199.42	0.20	8.41	0.16
Hyperactivist	921.81	19.61	2491.63	362.46	0.32	7.62	0.18

APPENDIX B: RESULTS OF THE CLUSTERING ANALYSIS OF THE TEN TIME WINDOWS IN SUPER USER

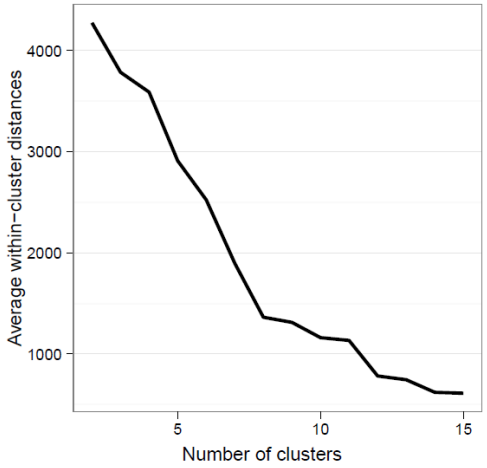

Figure 9: Analysis of the heterogeneity in the iterative cluster solutions in the hierarchical method applied to the ten time windows in Super User.

Table 5: Unstandardized cluster centers in the profiles identified for the ten time windows in Super User.

Profile	Answers	Questions	Comments	Activity Duration	MUAnswers	MUQuestions	MUComments
Low-activity	0.35	0.39	0.55	1.06	0.01	0.06	0.00
Occasional	0.65	1.72	2.66	2.70	-0.02	2.66	0.00
Unskilled answerer	1.00	0.09	0.33	1.13	-0.68	0.00	0.00
Unskilled commenter	1.13	0.30	1.78	2.06	-0.05	0.20	-0.57
Expert answerer	1.11	0.17	0.94	1.62	0.96	0.13	0.00
Expert questioner	0.98	0.88	2.81	2.11	0.03	12.90	0.01
Expert commenter	1.12	0.26	1.64	2.18	0.06	0.19	1.02
Q-A activist	4.36	11.88	18.23	12.70	-0.12	1.55	-0.04
Answer activist	33.58	1.42	41.79	23.43	0.10	1.18	0.09
Hyperactivist	166.81	2.06	238.12	44.02	0.16	2.74	0.09

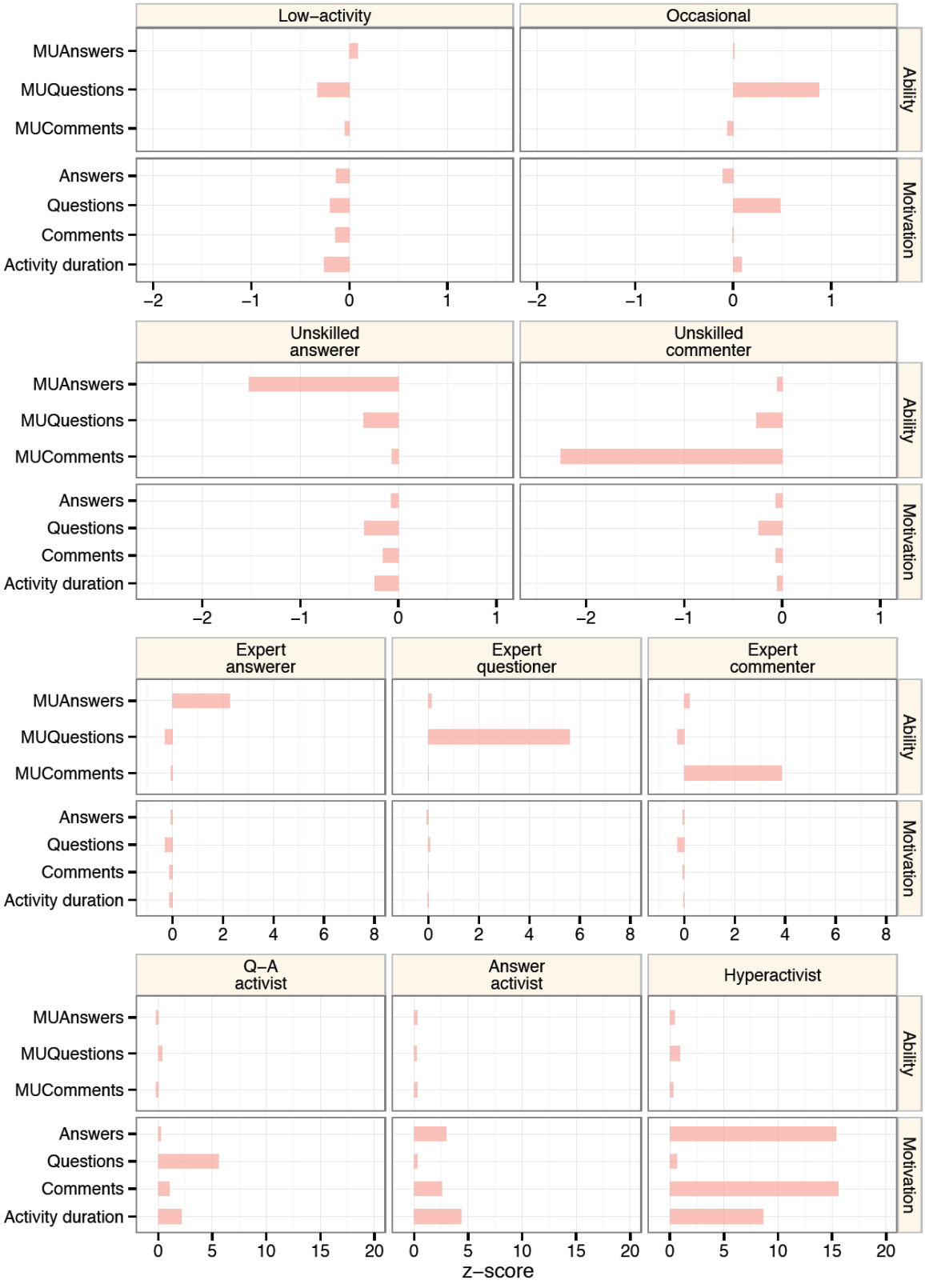

Figure 10: Centers of the clusters identified in the ten time windows for Super User.