Superposter behavior in MOOC forums

Jonathan Huang
Stanford University
jhuang11@stanford.edu

Anirban Dasgupta
Yahoo! Labs
anirban.dasgupta@gmail.com

Arpita Ghosh
Cornell University
arpitaghosh@cornell.edu

Jane Manning
Stanford University
jinpa@stanford.edu

Marc Sanders
Stanford University
sandersm@stanford.edu

ABSTRACT

Discussion forums, employed by MOOC providers as the primary mode of interaction among instructors and students, have emerged as one of the important components of online courses. We empirically study contribution behavior in these online collaborative learning forums using data from 44 MOOCs hosted on Coursera, focusing primarily on the highest-volume contributors—"superposters"—in a forum. We explore who these superposters are and study their engagement patterns across the MOOC platform, with a focus on the following question—to what extent is superposting a positive phenomenon for the forum? Specifically, while superposters clearly contribute heavily to the forum in terms of *quantity*, how do these contributions rate in terms of quality, and does this prolific posting behavior negatively impact contribution from the large remainder of students in the class?

We analyze these questions across the courses in our dataset, and find that superposters display above-average engagement across Coursera, enrolling in more courses and obtaining better grades than the average forum participant; additionally, students who are superposters in one course are significantly more likely to be superposters in other courses they take. In terms of utility, our analysis indicates that while being neither the fastest nor the most upvoted, superposters' responses are speedier and receive more upvotes than the average forum user's posts; a manual assessment of quality on a subset of this content supports this conclusion that a large fraction of superposter contributions indeed constitute useful content. Finally, we find that superposters' prolific contribution behavior does not 'drown out the silent majority'—high superposter activity correlates positively and significantly with higher overall activity and forum health, as measured by total contribution volume, higher average perceived utility in terms of received votes, and a smaller fraction of orphaned threads.

Author Keywords

massive open online course; MOOC; education; Coursera; collaborative learning; online forums; data mining

L@S'14, March 4–5, 2014, Atlanta, Georgia, USA.

http://dx.doi.org/10.1145/2556325.2566249

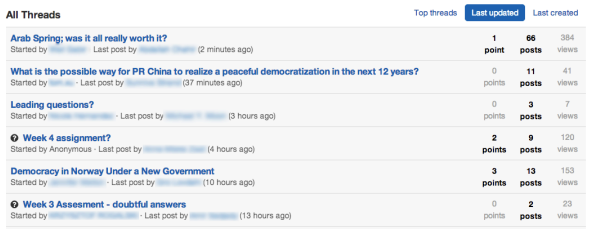
INTRODUCTION

Massive open online courses (MOOCs) have generated much excitement and interest because of their potential to bring about dramatic changes in higher education [11]. However, the very characteristics that enable the scalability of a MOOC—a handful of instructors using the Internet to broadcast lectures and content to a potentially unbounded number of students at once—have also engendered criticisms about the pedagogical soundness of this new model of engagement. It is not uncommon for the ratio of enrolled students to the number of teaching staff to exceed 5000:1, making it impossible for the large majority of students to have any meaningful interaction with the instructor. This lack of instructor attention is aggravated by the absence of an immediate peer group, which, in a physical classroom setting, is known to facilitate learning and understanding through discussion and tutoring [15, 3].

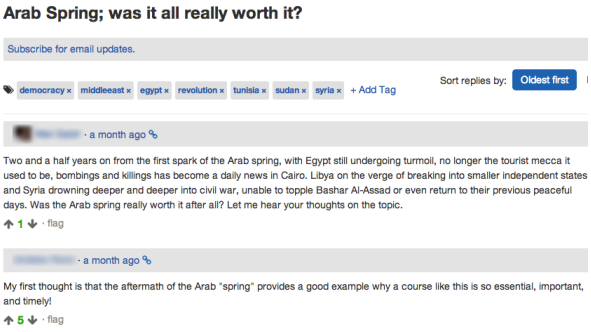
Forums, employed by MOOC providers as the primary mode of interaction among instructors and participants, have thus emerged as one of the critical components of a MOOC. Instructors use the forum to communicate about recent lectures or homework assignments, and have even been known to use structured open-ended questions on the forum to encourage discussions. Students express their views, seek help from peers and discuss assignments. It has been suggested that a well-run discussion forum provides a sense of community and engagement that is all but able to substitute for the peer support available in a physical classroom. Indeed, there is anecdotal evidence[1] that some users have found the active forums in particular MOOCs to be among the most important enablers in successfully completing their course.

In this paper, we study the most vocal subset of contributors on MOOC forums. We call these students the *superposters*—the students who post most frequently on the forum[2], and typically disproportionately more often than their peers[3]. Superposters exist in every course, and as the "loudest" participants, can play an outsized role in the quality and tone of discussion in MOOC forums. For a course designer, understanding superposters—their characteristics, behavior, and overall utility to forums—is central to understanding how (and

[1] `http://mooc.studentadvisor.com/posts/23/four-ways-to-get-the-most-out-of-a-mooc`
[2] `https://signalblog.stanford.edu/how-widely-used-are-mooc-forums-a-first-look/`
[3] This pattern is not at all uncommon in online environments, and specifically in online content production; see the section on related work.



(a)



(b)

Figure 1. (a) Screenshot of part of the front page of the discussion forum for Stanford's Democratic Development course; (b) Example thread from the same course.

whether) to encourage superposters, to identify roles such users could undertake in the forum that would not be feasible at scale for an instructor, and what might act as effective incentives for steering their contributions.

Superposters can, ideally, be model participants, making a large volume of timely high-quality contributions, and inspiring their peers by example to participate regularly in course discussions. But it is not, a priori, obvious that this ideal holds, or even that superposters are actually 'good' for forums— it is conceivable, for instance, that the superposters in a course could be flooding the forums with low quality posts, engaging in "trolling" behavior, or alienating the "silent majority" of the remaining students in the class. We therefore investigate the following questions in this paper. First, who are these superposters—what are their demographics and characteristics, and how do they engage across the MOOC platform, both in terms of forum behavior and course performance? Second, how 'useful' are the contributions from these high-volume users, and how does their activity correlate with contributions from other forum participants—that is, how do superposters' contributions relate to the utility of the forum as a whole?

Our contributions

In this paper, we embark on an investigation of superposters in MOOC discussion forums, using data from 44 courses hosted on Coursera during 2012-13. The fact that our dataset spans multiple instances of forums allows us a unique opportunity to go beyond studying contribution in a single instance of a collaborative learning forum, with all its associated restrictions, and investigate users' contribution patterns as well as overall forum health across multiple forums.

We first investigate superposters—their demographics, and contribution patterns such as length of posts and asking versus responding tendencies as in past research [8, 1]—as well as their engagement patterns across the Coursera platform. We find that superposters display above-average engagement and performance on Coursera, enrolling in more courses and obtaining better final grades than non-superposters. Most interestingly, we see that users who are superposters in one course are significantly more likely to be superposters in other courses they take as well, suggesting that superposting might be an inherent, individual-specific trait, rather than an extrinsically induced response arising from course or forum-specific circumstances.

We next address whether superposters contribute value, beyond volume, to the forums. Our data analysis indicates that while neither the fastest nor the most upvoted, superposters do respond faster and write longer posts than the average contributor. In addition, an assessment of quality on a subset of human-coded[4] superposter posts agrees with the inferences from the quantitative analysis, indicating that a large fraction of superposter contributions indeed constitute useful content. Finally, we explore whether superposters' prolific posting might negatively impact others' inclinations to contribute. An analysis of the correlation between superposters' activity levels and overall forum activity shows that high superposter activity correlates positively and significantly with higher overall activity and forum health, in terms of total contribution volume, received upvotes, and the number of orphaned threads.[5] Our results suggest that rather than flooding forums with low quality noise or 'drowning out the silent majority', superposters are, in some sense, 'model' forum citizens—users who contribute significant value to the forums through their effort, often across multiple courses.

A caveat emptor goes with our results. Our study is based purely on observational data collected "in the wild" without the explicit intent to perform research, so that our results are purely correlational—specifically, we do not claim causal conclusions from our analysis, nor make claims about learning outcomes which cannot be measured due to the nature of our data. A more rigorous experimental study, possibly with hypotheses informed by our results, that identifies causal effects of the behavior of high-volume contributors while controlling for possible confounding factors, can potentially provide useful input to platform designers in a number of ways. Such a study might be informative regarding how to engage and reward such users, what roles they can potentially take on to ensure scalability, and whether and how resources spent on identifying and 'incentivizing' superposters in one course may pay off across multiple courses on a MOOC platform.

Related work

Online forums for education, often referred to as asynchronous discussion groups, have been extensively studied in the computer-supported collaborative learning (CSCL) literature [8]. Collaborative learning is critical in education: by allowing learners to confront tasks or learn concepts that they

[4]While most content analyses in the CSCL literature have relied on manually coded data as in our own paper [5], it seems clear that automated natural language processing methods (such as those described in [12]) will be more scalable for future analyses.

[5]i.e., threads which receive no responses

would not be able to do alone, collaborative environments effectively form a *scaffolding* [20] for learners to proceed to their next developmental level [17].

Online collaborative learning interactions have been shown to enhance academic discourse, to foster higher level cognition of concepts [4, 9, 14], and to lead to gains in learning outcomes [14]. Additionally, asynchronous online discussions can play an important motivational role through promoting social presence and belonging [13, 18], and thus even messages that are seemingly "off-task" may potentially play a useful role in forum utility. Finally, the benefits of online collaborative learning extend beyond active participants to passive viewers too [6]. By using self-reported data from students, Dennen et al. [6, 16] report that students benefit from the forum content through a process of "reading and reflection".

While the majority of CSCL work has focused on smaller-scale studies, a number of large scale data analyses have been conducted on general-purpose online Q&A forums, such as Y! Answers and StackOverflow, in the data mining community. The study of contributors in these Q&A forums has largely focused on the question of identifying "experts", in order to identify high-quality content and curate or highlight contributions from such experts (see [10, 22, 1, 2] and references therein). In contrast, rather than identifying a subset of experts, we investigate the expertise (and other characteristics) of a given subset of contributors. The work of Furtado et al. [7] is most closely related to ours from this literature, undertaking a more general study of online contribution behavior by clustering and classifying contributors' activity profiles in Stack Overflow and Yahoo Answers. Their categorization of user contribution patterns yields a number of "activist" profiles—users who contribute a large number of questions and answers, although with answering skills that are only slightly better than the average. While our definition of superposters is simpler, we observe similar patterns of 'skill' for superposters in our analysis.

Finally, we note that the phenomenon of superposting behavior is not new to our study— the existence of a small number of core users who contribute disproportionately to the total content volume, resulting in the ubiquitous heavy tail in contribution sizes, is well-documented in multiple large-scale online communities [21, 19]. However, the contribution characteristics of such high-volume contributors have not been studied at scale in a collaborative learning context.

DATASET AND FORUM ORGANIZATION

We consider data from 44 courses run on Coursera with over 70,000 discussion threads spanning a range of topics, mostly from STEM disciplines (8 of the courses were non-STEM). Table 1 summarizes some of the statistics of our data. We note that with the exception of a handful of courses, forum participation is voluntary in Coursera courses.

The discussion forums on Coursera are organized as follows. Each forum is a 2-level tree of "sub-forums", typically with several subforums dedicated to general topics such as course logistics, errata, technical support, and study groups, as well as subforums for more course-specific topics (e.g., a "Queries" sub-forum may have further sub-forums such as "Unit 6 Queries"). Threads are also up to 2 levels deep, consisting of an ordered sequence of posts with additional comments optionally attached to some posts. We do not distinguish in this paper between comments and other types of posts, simply using the term "posts" for both. Students can vote posts or comments up or down (once per user per post), and are encouraged to use their votes to "bring attention to thoughtful, helpful posts" rather than express subjective agreement or disagreement.

# of MOOC offerings	44
Total # of threads	70,419
Total # of contributions (posts and comments)	325,071
Total # unique forum contributors over courses	116,028
Median # registered students per course	40,674
Median # of unique forum contributors per course	2180.5
Median # of threads per course	1,297

Table 1. Summary statistics of the datasets considered in this paper.

Students can choose to view only the threads from a particular sub-forum or browse through the most recent threads that have been posted anywhere in the forum. When browsing the list of threads, students can see the length of each thread, the number of times this thread has been viewed, the net number of upvotes this thread has received, and whether a staff member contributed to it. Students can view the contents of threads either chronologically or by popularity as determined by net votes per post. Finally, students can also subscribe to threads (and are subscribed by default to threads to which they contribute).

SUPERPOSTER CHARACTERIZATIONS

There are several measures that might be used to characterize the extent of a user's contribution to a forum. In this paper, we focus on the *quantity* and *'quality'* of a user's contributions. We will be particularly interested in definitions of superposting behavior that permit cross-course comparisons and analysis, which can be nontrivial due to different course durations; our measures and definitions are chosen accordingly to allow such comparisons.

We define *quantity* as the *number* of posts made by a student on course forums. (Note, however, that other natural measures for quantity, such as word counts, are also defensible in this context.) To account for different course durations, we define a user's *quantity score* for a course to be the average number of contributions she makes per week in that course. We define *(quantity) superposters* in a course to be the set of users who belong to the top 5% of forum participants in the course with respect to the quantity score. We note here that while we used a *relative* measure of contribution to define superposting behavior, there are a number of possible alternative definitions, including measures based on thresholds for the absolute number of posts, or the ratio between the number of posts to an average; we discuss and analyze these alternative definitions (for all three kinds of superposters defined in this section) in the full version of the paper.

Quality is a more elusive trait. While we would ideally like to measure to what extent a user was able to accurately and clearly answer questions and contribute fruitfully to discussions, estimating such a measure is not easy since rating the

accuracy of a forum post might, in many cases, require specialized domain knowledge about the course content. As a proxy, therefore, we use votes cast by other students in the course forum as an approximate measure of quality.

We define the quality of a user's contributions to a particular thread to be the ratio of the number of votes on all her contributions to the thread to the average number of votes on any contribution in this thread. A user's *quality score* in a course is the average of her per-thread quality over all the threads that she contributes to in the course. We say that a student is a *(quality) superposter* in a course if she is in the top 5% of forum participants in that course according to this quality score. Since a contributor who has a single highly-rated contribution could be propelled to being a quality superposter with this definition, we additionally set the quality score of a user to zero if they contributed fewer than five posts or comments throughout the course.

Coursera also maintains and displays a *reputation score* for each student, computed as the sum of square roots of votes across all contributions by a user. Reputation scores can be thought of as measuring both the quantity and quality of a user's contribution, while reducing the effect of votes on any single contribution. We again say that a student is a (reputation) superposter in a course if she belongs to the top 5% in the course with respect to the reputation score.

For each kind of superposter defined above, we use non-superposters to refer to the forum participants who are not superposters of that kind (note that we only consider the set of *forum participants*, rather than the entire population of registered students in a course, to define non-superposters). In this paper, we will primarily investigate the behavior of 'quantity' superposters—the users who contribute the largest volume of content on a forum; unless otherwise specified, the term *superposters* will henceforth refer to quantity superposters. While these (quantity) superposters are the main focus of our study, we also compare their behavior to the quality and reputation superposters to calibrate our observations wherever appropriate.

DEMOGRAPHICS AND PARTICIPATION PATTERNS

Who are these superposters, and what are their engagement patterns across the MOOC platform? In this section, we investigate superposter demographics, contribution patterns across course forums, and course enrollment and performance.

We begin with demographics, correlating superposting behavior with data from a survey conducted by Coursera that was administered by a large fraction (roughly two-thirds) of the courses in our dataset. Among these courses, 7% of the entire set of users, 17% of the forum posters and 100% of the superposters filled in the survey. While it may not be surprising that native English speakers tend to be the more vocal participants on Coursera forums, other demographic factors also play a role— for instance, the histogram of forum participants by age in Figure 2(a), partitioned into superposters and non-superposters, shows that superposters are typically older than the average forum user (and Coursera users in general). Gender also plays a small but statistically significant

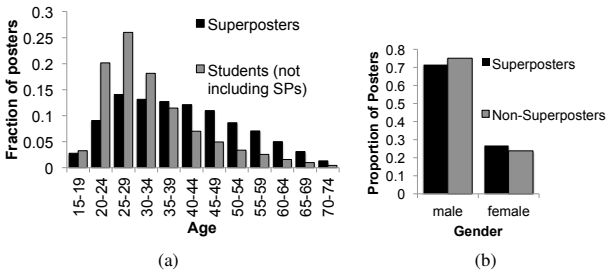


Figure 2. Demographic histograms of age (a) and gender (b) from Coursera survey, comparing superposters and non-superposters.

role—while the majority of forum participants (as well as superposters) are male, the proportion of superposters who are female is slightly higher than that of non-superposters.

Contribution characteristics

We now explore two basic characteristics of the contributions superposters make on MOOC forums. Previous work has studied the overlap of 'askers' and 'answerers' in online Q&A forums: for instance, Adamic et al. [1] find that Yahoo! Answers has subforums with a significant fraction of users who both ask and answer, as well as subforums where users almost exclusively either ask or answer questions. Analogously, we would like to understand whether superposters achieve their large posting volume primarily by 'asking' or 'answering' questions, or a mix of both. Unlike in many such non-educational Q&A forums, however, the first post of a thread in Coursera forums does not necessarily have to be a question (although it typically is one); also, nothing prevents a student from asking a question midway through a thread, prompted by the preceding discussion. Thus, instead of distinguishing between asking and answering a question, we distinguish between *initiating* and *responding* to a thread, and ask whether superposters in MOOC forums tend to be initiators or responders.

The tables below list the number of initial posts and responses (i.e., posts that are not the initial post in a thread) from superposters and non-superposters. These numbers show that the ratio of the number of responses to the number of threads initiated is greater for superposters than for non-superposters (by almost 4 responses for each thread initiated), suggesting that superposters tend to respond to threads (possibly by answering questions) more often than starting a new thread (possibly by asking a question). Next we study the length

(a) Superposters

# responses:	208,690
# threads initiated:	20,629
# responses to # threads initiated ratio:	10.12

(b) Non-superposters

# responses:	265,566
# threads initiated:	42,545
# responses to # threads initiated ratio:	6.24

of posts, which can be viewed as a proxy for quality as well as an alternative measure of quantity. Measuring the number of posts from a user might not actually reflect the volume of her contributions to the forum if users trade off post length and quantity, with some users writing many short posts but

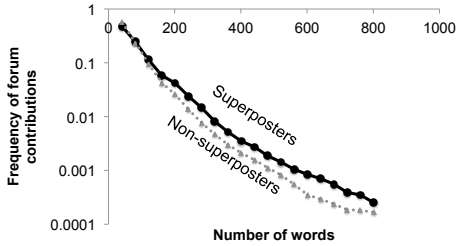


Figure 3. Histogram of contribution lengths (number of words in a post or comment) comparing superposters and non-superposters. In addition to writing more posts, superposter contributions tend to be longer.

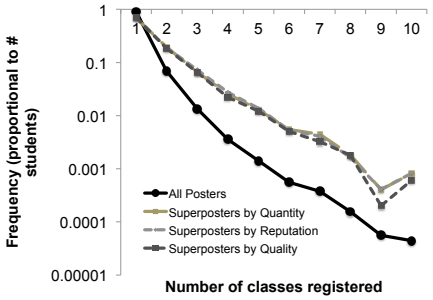


Figure 4. Histogram of the number of courses to which each forum poster has enrolled (of the courses in our dataset), showing that superposters are typically enrolled in more courses than non-superposters. The y-axis in this plot is probability plotted on a log-scale.

contributing the same 'total volume' of forum conversation as users who write a handful of long posts. Figure 3 suggests, however, that user behavior on the forum does not display such a "conservation of words" effect: the histogram of the number of words per post in contributions from superposters and non-superposters in Figure 3 shows that in addition to responding more frequently, superposters are also more likely than non-superposters to write lengthier posts.

Superposting across courses

Is superposting behavior an inherent trait, where some posters simply are prolific irrespective of the environment while others are not, or is it driven by extrinsic factors such as course content or the forum environment and community?

We approach the question of whether superposting behavior is an inherent or extrinsic trait by focusing on users who were enrolled in multiple courses in our dataset. While most forum participants were only enrolled in a single course, roughly 8000 forum posters in our dataset were enrolled in more than one course. Of these, we focus on the $\sim$ 6200 students enrolled in exactly two courses; 900 of whom were superposters in at least one course. Table 2 tallies the number of these students who were (1) superposters in neither of the two courses, (2) in exactly one course, or (3) in both courses. As the table shows, if a student was a superposter in one class, she was significantly more likely (nearly three times more likely than would be expected under independence) to be a superposter in another course. Testing the null hypothesis that superposting behavior for an individual is independent across courses with a chi-squared test run on Table 2, we can reject independence (with $\chi^2 = 205.34$, $p \leq .01$). Furthermore, this pattern holds for each of our three definitions of superposters (i.e., by reputation or quality). We therefore conclude that superposting behavior is persistent across multiple courses and appears, at least to some extent, to be an inherent trait. A more careful study of this phenomenon, as well as a further understanding of superposter motivation, can potentially have implications for platform design since identifying or incentivizing superposters in one course may yield payoffs across multiple courses.

	Non-SP in Course 2	SP in Course 2
Non-SP in Course 1	5328	386
SP in Course 1	386	128

Table 2. Contingency table counting number of students who were superposters in zero, one or two courses.

Course enrollment and performance

Finally, we investigate how superposters engage across the Coursera platform—are superposters more engaged with MOOCs overall, and do they do well in courses or are they the weaker students in the class, coming repeatedly to the forum for assistance? We begin by investigating superposter course enrollment via a histogram of the number of courses in which superposters and non-superposters are enrolled. Comparing non-superposter enrollment rates to that of superposters in Figure 4, we see that superposters have a tendency to enroll in more courses on Coursera than 'regular' students.

We next investigate course performance as measured by grades. Our analysis suggests that on average, superposters tend to also be the better performers in courses, although the extent to which they outperform other students depends on subject matter. Before describing this analysis, we note that this (purely correlational) result suggests an immediate open question regarding the direction of causality between forum participation and learning outcomes: while one plausible hypothesis is that high expertise (likely leading to good performance), gives students the confidence to be vocal on forums, an alternative hypothesis is that high forum participation leads to good performance because asking questions and explaining material to others leads to better learning outcomes. Our (observational) data cannot properly address this question; however, resolving this question via an experimental study is an important direction for further work.

To evaluate the course performance of superposters, we compute the average z-score of the final course grades of superposters and non-superposters and examine the difference between these averages. This *grade disparity* indicates the number of standard deviations by which superposters outperformed other students on average. A potential confounding factor is time of engagement—a student who was actively engaged for only 4 weeks out of an 8-week course is unlikely to have obtained a good grade, and even less likely to have had among the highest average rates of forum posts per week. To address this, we use the fraction of lectures opened by a user as a proxy for time of engagement, and control for time of engagement by including only those users who accessed sufficiently many lectures (we filter out any user who opened fewer than 10% of the lectures in a course).

Figure 5(a) plots the results of this analysis with respect to all three of our superposter definitions. In each case, we see that superposters outperform their peers by approximately

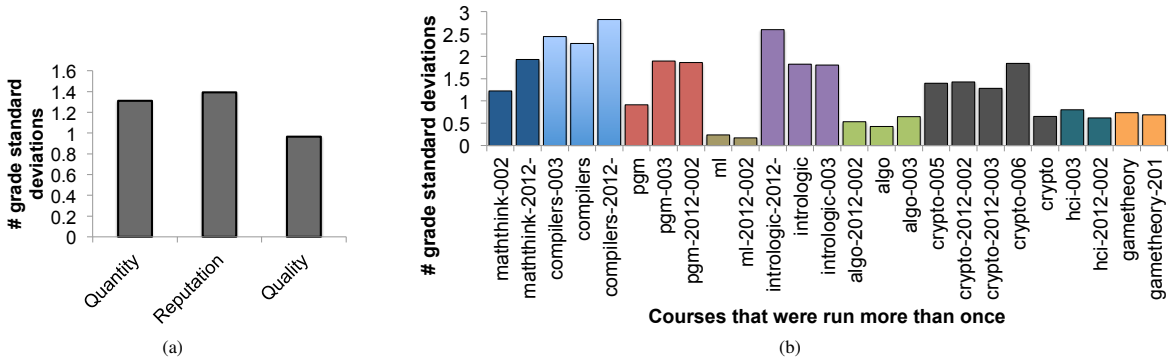


Figure 5. (a) # of standard deviations by which superposters outperformed non-superposters on final course grade (controlling for time of engagement), averaged over all courses, with respect to all three definitions of superposters. (b) Per-class results of for quantity superposters on the same data, shown just on courses in our dataset which were run more than once (best viewed in color, with colors indicating multiple runs of the same course)

one standard deviation, after controlling for time of engagement. Superposters by quantity outperform superposters by quality, while superposters by reputation perform best—that is, users whose forum posts balance quantity and quality (per our vote-based metric) tend to also be the best students.

Examining the grade improvements on a course-by-course basis yields a further insight. Figure 5(b) plots the grade disparity for nine courses which were run multiple times. In each case, we see that superposters outperformed their peers, on average, and the amount of grade disparity varied from course to course; notably, though, the grade disparity is similar across multiple offerings of the same course. While more work is needed to understand these similarities, reasonable hypotheses might be that the level of confidence required to be a superposter and/or the learning gain from posting at volume as a superposter is subject-dependent.

SUPERPOSTERS AND VALUE CREATION

In this section, we investigate whether superposters create value—beyond quantity—in the forums. As such, value is a fairly broad concept, and there are a number of metrics that might be used to measure the value, or utility, of a contribution. We will use two natural and easily computed quantities, namely posts' response times and received votes, to reflect two important aspects of healthy forums—whether questions are answered quickly without much delay, and whether other forum users react positively to contributions.

Response times

We use two different measures for response time. Our first measure is the absolute time to respond, i.e., the difference in time between the initial post in the thread and the time at which a contributor posts a response. Figure 6 shows the histogram of superposter response times, where a user's response time is the average, over all threads to which the user contributed, of the time to her first response within the thread. For comparison, we also plot the histograms of the absolute average response time for three other categories of users: the *earliest responders*, the quality superposters, and the reputation superposters. As with superposting, we define *early responders* as the set of users who have answered at least five questions, and belong to the fastest 5% of users by average response time, computed over all the threads that a user

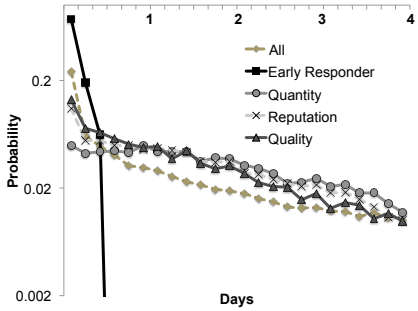


Figure 6. Histogram of response times (defined as the amount of time elapsed from the time a thread is started to time of post or comment), comparing early responders, superposters and the background distribution of all forum participants.

responded to. We also include the response time histogram computed over all users in all forums.

This analysis shows that while having better-than-average response times, superposters are not the quickest responders in a forum: early responders post responses on threads within the first 12 hours if at all (with a median response time of 6 hours), while an average forum user posts a response within 2 days of the question being posed 98% of the time (with a median response time of 43 hours). Superposter response times more closely resemble those of typical forum users, with median response times of 56 hours for quantity superposters, 45 hours for reputation superposters, and 36 hours for quality superposters.

A different measure of response time is ordinal, rather than cardinal as in the previous plot—what is the relative rank (in order of arrival) of a superposter's response among all responses in a thread? Our second measure is based on the order of arrival of posts in a thread rather than the absolute time to respond: for each thread, we divide all responses (excluding the first question or post that started the thread) into quartiles, ignoring threads with fewer than 5 posts. For each category of superposters (quantity, reputation and quality), we then count the fraction of posts from that category of users lying in each quartile.

Figures 7(a) and (b) show the histograms for the quartile position for each user category (superposters and non-

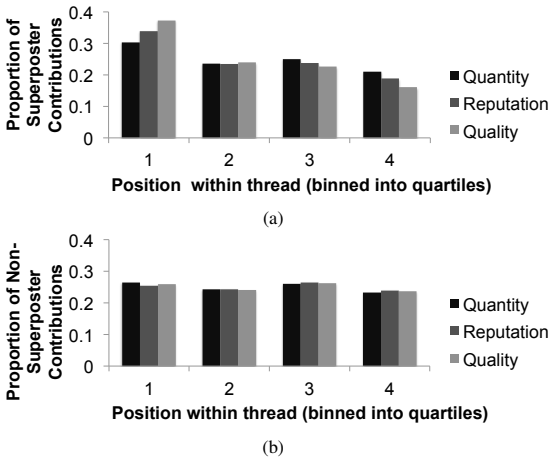
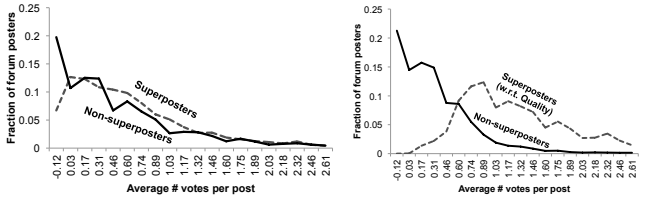


Figure 7. Histograms of the positions within a thread at which superposters (a) or non-superposters (b) contributed. We bin the positions coarsely into quarters (first quarter of thread, second quarter, etc.), ignoring threads with less than five posts and not counting the first post (i.e. the question that started a thread).



(a) Superposters by quantity (b) Superposters by quality

Figure 8. (a) Histogram of votes per post over forum participants, comparing superposters to non-superposters; (b) The same histogram with respect to the quality superposter category.

superposters by quantity, quality and reputation). As expected, the average user's post is equally likely to belong to any of the four quartiles. On the other hand, we see that superposter posts are slightly more likely to be in the first quartile than average (0.30 instead of 0.25); also, superposters are more likely to post responses that fall in the first quartile than in any other quartile. Thus, superposters, although not the earliest of responders, typically post responses earlier in the thread than an average responder would. (We note here that the observation from the histogram that quality superposters are more likely to respond in the first quartile (0.37 fraction of the time) than the two other kinds of superposters might be due to quick responses receiving more upvotes, and therefore increasing a user's quality score in our definition.)

Votes

As a second measure of contribution value, we analyze the votes received by superposters' contributions. While votes, in general, need not indicate the accuracy or completeness or any other absolute measure of a post's quality as previously discussed, an upvote from a user on a contribution does indicate a degree of satisfaction or utility that was derived from that contribution (for whatever reason). Thus votes arguably do provide a reasonable indication of the usefulness of a contribution.

To measure the value of a user's contributions from voting data, we take the average of the votes (both up and down) received over all posts or comments made by a user in a course. Specifically, for each user in a course forum who contributed at least five times, we compute the mean of the number of votes given to each of her contributions by other users. In Figure 8(a), we plot the histogram over these computed means for quantity superposters and non-superposters, discarding the part of the histogram outside the 5th or the 95th percentile to remove outliers. (For comparison, Figure 8(b) shows analogous histograms using quality contributors.)

We see that the histogram of votes on posts by superposters, while somewhat better, is not very much higher than that for an average user, suggesting that superposters do not (at least consistently) produce posts that other users upvote significantly more than those from non-superposters. The median number of votes per post for a superposter is 0.56, while the median vote is 0.4 for a non-superposter, which is about 28% smaller. Therefore, while superposters do produce better-than average quality content (as measured by votes), superposters are outstanding more for the quantity than for the quality of their contributions to the forum.

A closer look at superposter content

We now supplement our quantitative analysis of the quality of superposters' contributions with a manual assessment of a subset of superposter posts. Instead of performing an exhaustive or comprehensive qualitative analysis, we ask what percentage of superposter posts in this subset could be described as content-focused and positive—*on-content posts*. We manually examined the posts and comments from the top 3 superposters in each of 4 classes (a total of 1996 posts in all), and classified them simply as being on-content or off-content. Coding was performed independently by two of the authors. We remark here that contributions that were not on-content were not necessarily "bad" — some were simply phatic or logistical in nature. Posts categorized as on-content included ones that answered or asked a content-related question, engaged in content-related dialog in a productive way, or directed people to related resources.

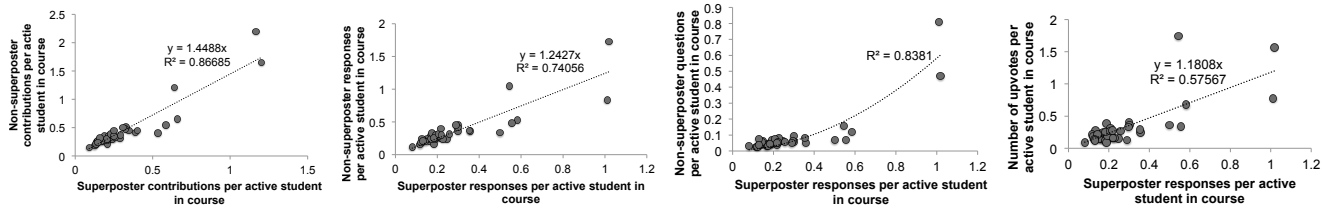
Our findings are summarized in Table 3. Overall, 68.8% of posts from the top 3 superposters in a course were rated as on-content, which is a fairly high fraction given the stringent definition we used. One way to interpret these findings is to view them as support for the notion that online discussion forums for these courses effectively mimic face-to-face study sessions: students ask and answer questions about the course content, and while there is some chatter (occasionally regarding course logistics and at other times simply extraneous to the course), the content coming from the most visible students models the behavior we hope to see in any kind of study group, whether online or in-person.

SUPERPOSTERS AND OVERALL FORUM ACTIVITY

The previous section studied superposters' direct effect on forum health, asking whether superposters' contributions bring value, as measured by upvotes and the speed of response, to the forum. But even if superposters do contribute a large quantity of content of reasonable quality, they might still not be an entirely positive influence on the forum if their

Course	# posts from top 3 superposters	% on-content posts from top 3 superposters	% on-content posts from superposter 1	% on-content posts from superposter 2	% on-content posts from superposter 3
Child Nutrition	521	81%	79%	95%	77%
Algorithms	380	80%	80%	71%	90%
Intro To Logic	551	79%	88%	83%	47%
Writing in the Sciences	544	39%	32%	22%	60%

Table 3. Summary of results of qualitative study of contributions from the top 3 superposters from four representative courses, in which posts and comments were manually coded as on or off-content.



(a) (b) (c) (d)

Figure 9. Scatterplots measuring superposter (SP) influence on non-superposter (nonSP) forum behavior (with each point corresponding to a course): (a) # SP contributions vs. # nonSP contributions; (b) # SP responses vs. # nonSP responses; (c) # SP responses vs. # nonSP threads initiated; (d) # responses vs. # votes obtained by nonSPs.

prolific posting suppresses contribution (for a variety of plausible reasons) from the remainder of the class. In this section, we investigate the correlations between superposter contributions and overall forum activity.

Relationship with overall participation and quality

We first study how superposter contribution relates to participation and contribution quality from other users in the forum.

We begin with quantity. Figure 9(a) shows a scatter plot of the number of posts by non-superposters in a forum against the number of superposter posts, where each point corresponds to a single course, and both axes are normalized by the number of *active students* in the course (a student is defined as active if she opened at least 5% of the course lectures). The line of best fit in the scatter plot shows a high positive correlation ($R^2 = 0.86$) between the two quantities, suggesting that higher activity from superposters is positively correlated with higher activity from other forum users as well[6]. Of course, this effect need not be causal at all, and could arise entirely from some latent factor (such as instructor encouragement or incentives for participation) that leads to high activity by all forum users—that is, we do not claim that superposter activity begets more activity from non-superposters; however, the analysis does suggest that superposter activity does not *suppress* non-superposter activity on the forums.

We next examine the correlation between superposter posts and non-superposter contributions in greater detail. Figure 9(b) plots the number of *responses* (i.e., excluding first posts) by non-superposters against the number of superposter responses, again normalizing both quantities by the number of active students in the class. Figure 9(c) has the same x-axis as Figure 9(b), but the y-axis represents the number of threads *initiated* by non-superposters. Figure 9(b) shows a linear correlation ($R^2 = 0.75$) between the quantities whereas a quadratic fit was more appropriate for Figure 9(c). Again, these plots all suggest that high activity from superposters does not negatively impact participation—either initiation or response—from non-superposters.

Finally, we study the correlation between superposter activity and the quality of non-superposter contributions. Figure 9(d) is a scatter plot with the same x-axis as Figures 9(b) and 9(c), and the average number of upvotes over non-superposters posts (normalized by the number of active users) on the y-axis. We again observe a linear correlation (with $R^2 = 0.576$) which, though not as a strong as in Figures 9(b) and 9(c), indicates that a larger number of superposter responses also correlates positively with an increase in the number of upvotes received by a non-superposter.

Relationship with number of orphaned threads

While most threads posted on discussion forums in the classes in our dataset receive responses, some fraction of threads go unanswered. Among courses where posting is not required, we find that between 10% and 50% of threads started are "orphaned", i.e., do not receive any responses; this percentage is higher in courses where posting is required.

We next study the correlation between superposter activity and the proportion of orphaned threads in a course. Figure 10(a) is a scatter plot of the fraction of posts by superposters (among all posts in a course forum) against the fraction of threads left orphaned in that course (for this plot, we have removed 3 outlier courses in which participation on the forum counted towards a student's grade in the course). The Pearson correlation between fraction of posts by superposters and fraction of threads left orphaned is -.36 (i.e., negative) with $p = .02$, suggesting that courses in which superposters contribute more in volume tend to have fewer orphaned threads.

The question of what mechanism causes this correlation between superposter volume and orphaned threads remains open: one might speculate that the fraction of orphaned

[6]Note that a positive correlation between the unnormalized (or absolute) number of posts from non-superposters and the unnormalized number of posts from superposters need not convey information about whether superposters suppress participation from other users, since such a negative correlation, if not large enough, may be drowned out by an upward scaling in absolute contribution level with class size. This is why we normalize both axes by (effective) class size in Figure 9.)

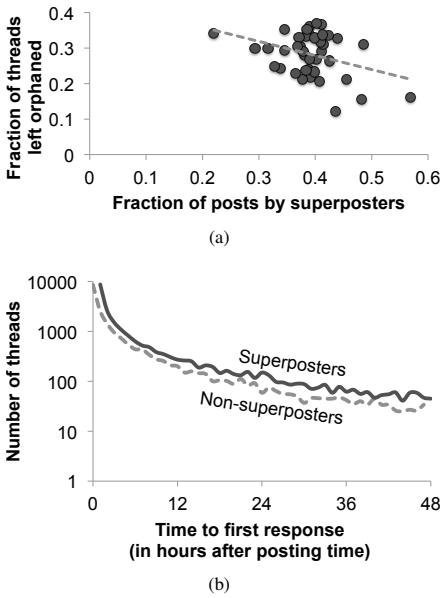


Figure 10. (a) Fraction of posts by superposters vs. fraction of threads left orphaned (each point correspond to a course); (b) Comparison of the number of threads receiving a first response from a superposter versus a non-superposter at k hours after posting time (for $k = 1, \ldots, 48$).

threads decreases with increased superposter activity because superposters are more likely to respond to threads that might otherwise have been left orphaned. However, Figure 10(b) compares, for all of the threads which got a first response in k hours, the number of first responses by superposters and non-superposters. The plot shows that even the threads that do not receive any responses for a long time (around 48 hours) after creation are no more likely to receive a response from a superposter than from a non-superposter, suggesting that the correlation between superposters and orphaned threads is more indirect, and possibly related to other course-specific factors. Like the other results in this section, these results are also perhaps best interpreted as the absence of a negative impact of superposting activity on orphaned threads (which is another measure of overall forum health) than an indication of any possibly positive causal effects resulting from superposters' contributions.

CONCLUSION

In this paper, we began an exploration of contribution patterns on MOOC discussion forums, studying user behavior and overall activity and forum health across 44 courses on Coursera. As in many large online communities, a large fraction of contributions in these collaborative learning forums come from a small subset of users, whom we refer to as 'superposters'. We investigate the characteristics of 'superposters'—their contribution patterns, demographics, course performance and enrollment—as well as the characteristics of their contributions—response speed, post length and quantity, perceived value as measured by upvotes (supplemented via human assessment on a subset of contributions), and finally how superposter activity correlates with participation from the rest of the class.

Our results suggest that superposting, which appears to be more an inherent than an extrinsic trait, largely results in high-value contributions and also correlates positively with activity and contribution quality from fellow students, mitigating concerns about contribution quality and any negative effects of such prolific posting on other forum users. Our study, being based on purely observational data, only allows drawing correlational rather than causal conclusions, and therefore suggests several immediate directions for experimental work, including (i) a further investigation of the possible 'inherentness' of superposting behavior and consequent implications for incentive design, (ii) dependencies between forum contribution patterns and the nature of the course content, and (iii) an experimental design to identify any causality in the correlation we observe between high forum contribution levels and strong course performance, with potential implications for improving educational outcomes and course design.

Our analysis, in addition to yielding insights about superposters and several hypotheses for further study, also yields a positive outlook on existing forums. While current forum designs are undoubtedly imperfect, with repetitive threads and rudimentary search and sort functionality with no interface for finding related questions, the forums do appear to provide reasonable utility. Participants used the forums for productive dialog about the class, ranging from quick questions and answers to sustained conversations, and the forums were mainly "healthy" in the MOOCs in our dataset—students who posted questions tended to get responses, and the students with the largest footprints participated in ways that were mainly positive, content-focused, and appreciated by other students. We note that by and large, these forums managed to thrive despite not adopting complex incentive schemes to encourage contributions, as in some other Q&A forums, leading immediately to the question of how much of forum contribution is driven by intrinsic, rather than extrinsic, motivation (this also relates to the 'inherentness' of superposting behavior that we observed in our analysis). However, it is worth noting that while these early MOOCs have had successful forums despite rudimentary design, maintaining healthy forums consistently over the long term might require adoption of design techniques—such as moderation privilege design or enabling direct acknowledgements of contribution—from existing successful long-running forums, such as StackExchange or Quora.

While forums for collaborative learning can improve student motivation and lead to learning gains, reaping these benefits in MOOCs relies on having healthy, active forums in a world of anonymity and little accountability, where participation is optional and casual lurking is the norm. A future in which a MOOC is run 100 or 1000 times (or simply left running continuously) is not unrealistic, and is one in which students may not have the luxury of instructors or TAs moderating forums. In such a future, a more thorough analysis of superposters— of their motivations, to understand how best to elicit high-quality contribution and sustained engagement from them—and of their abilities, to understand how best to utilize their efforts to create the most effective collaborative learning environments—may well become central to scaling the learning value and functions provided by MOOC forums.

ACKNOWLEDGMENTS

J. Huang is supported by an NSF CI Fellowship and acknowledges the support of Leo Guibas. We also thank Chris Manning for feedback on the paper.

REFERENCES

1. Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, ACM (New York, NY, USA, 2008), 665–674.
2. Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, ACM (New York, NY, USA, 2012), 850–858.
3. Crouch, C. H., and Mazur, E. Peer instruction: Ten years of experience and results. *American Journal of Physics 69* (2001), 970.
4. De Smet, M., Van Keer, H., and Valcke, M. Blending asynchronous discussion groups and peer tutoring in higher education: An exploratory study of online peer tutoring behaviour. *Computers & Education 50*, 1 (2008), 207–223.
5. De Wever, B., Schellens, T., Valcke, M., and Van Keer, H. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education 46*, 1 (2006), 6–28.
6. Dennen, V. P. Pedagogical lurking: Student engagement in non-posting discussion behavior. *Computers in Human Behavior 24*, 4 (2008), 1624–1633.
7. Furtado, A., Andrade, N., Oliveira, N., and Brasileiro, F. Contributor profiles, their dynamics, and their importance in five q&a sites. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, ACM (New York, NY, USA, 2013), 1237–1252.
8. Henri, F. Computer conferencing and content analysis. In *Collaborative learning through computer conferencing*. Springer, 1992, 117–136.
9. Ke, F., and Xie, K. Toward deep learning for adult students in online courses. *The Internet and Higher Education 12*, 3 (2009), 136–145.
10. Pal, A., Farzan, R., Konstan, J. A., and Kraut, R. E. Early detection of potential experts in question answering communities. In *User Modeling, Adaption and Personalization*. Springer, 2011, 231–242.
11. Pappano, L. The Year of the MOOC. New York Times, 2012.
12. Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., and Fischer, F. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning 3*, 3 (2008), 237–271.
13. Rourke, L., Anderson, T., Garrison, D. R., and Archer, W. Assessing social presence in asynchronous text-based computer conferencing. *The Journal of Distance Education/Revue de l'Éducation à Distance 14*, 2 (2007), 50–71.
14. Schellens, T., and Valcke, M. Fostering knowledge construction in university students through asynchronous discussion groups. *Computers & Education 46*, 4 (2006), 349–370.
15. Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., and Su, T. T. Why peer discussion improves student performance on in-class concept questions. *Science 323*, 5910 (2009), 122–124.
16. Soroka, V., and Rafaeli, S. Invisible participants: how cultural capital relates to lurking behavior. In *Proceedings of the 15th international conference on World Wide Web*, ACM (2006), 163–172.
17. Vygotski, L. S. *Mind in society: The development of higher psychological processes*. Harvard university press, 1978.
18. Walton, G. M., Cohen, G. L., Cwir, D., and Spencer, S. J. Mere belonging: The power of social connections. *Journal of personality and social psychology 102*, 3 (2012), 513.
19. Wilkinson, D. M. Strong regularities in online peer production. In *Proceedings of the 9th ACM conference on Electronic commerce*, ACM (2008), 302–309.
20. Wood, D., Bruner, J. S., and Ross, G. The role of tutoring in problem solving*. *Journal of child psychology and psychiatry 17*, 2 (1976), 89–100.
21. Wu, F., Wilkinson, D. M., and Huberman, B. A. Feedback loops of attention in peer production. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, IEEE Computer Society (Washington, DC, USA, 2009), 409–415.
22. Zhang, J., Ackerman, M. S., and Adamic, L. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, ACM (2007), 221–230.