Crowdsourcing with All-pay Auctions: a Field Experiment on Taskcn

Tracy Xiao Liu **Jiang Yang** **Lada A. Adamic** **Yan Chen**
University of Michigan School of Information
105 South State Street, Ann Arbor, MI 48109
liuxiao{yangjian,ladamic,yanchen}@umich.edu

ABSTRACT

We investigate the effects of various design features of all-pay auction crowdsourcing sites by conducting a field experiment on Taskcn, one of the largest crowdsourcing sites in China where all-pay auction mechanisms are used. Specifically, we study the effects of price, reserve price in the form of the early entry of high-quality answers (shill answers), and reputation systems on answer quantity and quality by posting translation and programming tasks on Taskcn. We find significant price effects on both the number of submissions and answer quality, and that tasks with shill answers have pronounced lower answer quality, which are consistent with our theoretical predictions. In addition, monetary incentives and the existence of shill answers have different effects on users with differing experience and expertise levels.

Keywords
Crowdsourcing, auction, field experiment.

BACKGROUND

One of the most remarkable and transformative potentials of the Internet lies in its ability to change people's collaborative work, especially in collecting intellectual contributions from disparate peer users on a large scale. This trend has manifested itself in various familiar examples such as open source projects, Wikipedia, Question-and-Answer (Q&A) forums, and social content and tagging sites such as Flickr, Del.icio.us and YouTube. In one type of collaboration called "crowdsourcing" tasks are directly outsourced to individual workers through public solicitation (Howe, 2006; Kleeman, Voss, & Rieder, 2008). Crowdsourcing sites have been rapidly growing in number, popularity, and research attention. For example, Taskcn.com, one of the earliest sites to have been studied, is a Chinese website where people post diverse tasks (e.g., design a company logo or translate a research statement) with a monetary reward for other users to compete for by submitting solutions (Yang, Adamic, & Ackerman, 2008). Amazon's Mechanical Turk is designed to invoke human labor to accomplish "human intelligence tasks" (HITs) requested by users with specified compensation (Mason & Watts, 2009).

Unlike many other kinds of "peer contributed" sites like Wikipedia or Flickr, crowdsourcing sites are task-driven with arbitrary requirements (or expectations) such as completion time, quality, or other features. These semi- or well-defined "tasks" might inspire less intrinsic motivation derived from some form of social reward (Nov, Naaman, & Ye, 2008) than free-structured and undefined contribution tasks. Thus, financial incentives have been increasingly incorporated into the design of crowdsourcing services. For example, Taskcn and Amazon's Mechanical Turk both allow requesters to set up monetary rewards in order to incentivize contributors. Q&A forums, such as Yahoo! Answers (Zhang, Adamic, Ackerman, & Bakshy, 2008), and the now-defunct Google Answers (Chen, Ho, & Kim, 2010) have employed varied incentive schemes to outsource knowledge or expertise requests. These schemes range from semi-market-like flat-rate virtual currency in Yahoo! Answers and virtual currency with a flexible rate as in Baidu Knows and Naver Knowledge to real-market-like Google Answers where real money is offered in exchange for knowledge or expertise.

Whether and how incentives can motivate more and better contributions have been the primary questions concerning economists and sociologists. Field experiments conducted on a series of Q&A sites have indicated that a higher reward can induce more answer submissions, but yield mixed results regarding answer quality (Chen et al. 2010; Harper, Raban, Rafaeli, & Konstan, 2008). Consistent results are found on Amazon's Mechanical Turk where financial incentives increase the quantity of contributions, but not quality (Mason & Watts, 2009). Similar results proving that higher awards elicit more answers are also found in field studies on Taskcn (Yang et al., 2008) and NaverKnowledge-In (Nam, Adamic, & Ackerman, 2009).

However, this does not paint a complete picture due to the inherent complications of differing types of required knowledge, tasks, incentive schemes, and communities. For example, although the amount of monetary award is

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

significantly correlated with the number of submissions on Taskcn, it could be confounded with the fact that people post higher prices for tasks that require a high level of expertise (Yang et al., 2008). A controlled field experiments for revealing how this crowdsourcing mechanism works in the context of the Internet knowledge market is therefore required.

THEORETICAL FRAMEWORK

We model the exchange mechanism on Taskcn as the all-pay auction in economics. Specifically, any user can submit an answer to a task and each task gets many different answers. Since every user who submits a solution expends effort regardless of whether or not they win, the knowledge-exchange mechanism is analogous to an all-pay auction where everyone pays for their bids in the form of individual effort, but only the winner gets paid.

Following Segev and Sela (2011), we characterize the sub-game perfect equilibria of incomplete information all-pay auctions. In particular, we predict how different monetary incentives and the existence of a reserve-price affect players' participation and effort level. In detail, we show (1) higher rewards induce more participation and more effort; (2) higher reserve price decreases the number of participation. Furthermore, there is an optimal reserve price which generates the highest effort. It implies that if the reserve price is too high, it will decrease individuals' effort.

The complete model and proof can be found on the corresponding author's website.[1]

EXPERIMENTAL DESIGN

We use a 2x3 factorial design to investigate the price and the reserve price effects on users' behaviors. Specifically, we are interested in understanding whether tasks with higher rewards would attract more submissions as well as higher answer quality. We are also interested in determining whether a high-quality answer posted early can deter the entry of late answers, especially if it is posted by a user with a history of winning on the site.

Task Selection: Programming and Translation Tasks

We choose to use translation and programming tasks for this experiment, as the quality of these two types of tasks is quite standard and objective.

For programming tasks, we collected 28 real programming problems from students at the University of Michigan School of Information, consisting of 14 Javascript and 14 Perl tasks. All these tasks were not searchable and had practical implications. In the experiment, they were randomly assigned to different price treatments. We were unable to provide shill answers for programming tasks due to their difficulty. Consequently, we only used translation tasks in the shill treatments.

For translation tasks, we selected two types of translation work: personal statements collected from Chinese graduate students at the University of Michigan and company introductions downloaded from Chinese websites. We chose these two types of translation tasks because they are challenging, requiring a high level of skill and effort compared to other types of translation work, such as translating a CV. For each translation in shill treatments, we provided a shill answer which was either provided by the personal statements' owners or created by two of our undergraduate research assistants. To ensure that the shill answer had a relatively high quality, we asked one Chinese student to translate each company introduction from Chinese to English, and then asked the other American student to revise it.

Treatments

To investigate the price effects, we choose two prices for our tasks: 100 Yuan and 300 Yuan. First, as 100 yuan, is the empirical median price for both programming and translation tasks, it guarantees a certain amount of participation, even for low-price treatments. Second, the gap between 100 and 300 is salient enough for us to observe price effects on users' behaviors. Altogether, we have six different treatments in this experiment:

1. High Price, No Shill: each task is posted with 300 Yuan as a reward.
2. High Price, Shill without credit: each task is posted with 300 Yuan as a reward. On average, within three hours after the task is posted, we post a shill answer. Each shill is posted under a different user's name.
3. High Price, Shill with credit: each task is posted with 300 Yuan as a reward. Averagely, within three hours after the task is posted, we post a shill answer using an existing account on Taskcn. The owner of this account has 4 credits, representing a relatively high winning record on the site. Users earn 1 credit whenever they earn 100 yuan on the site. We developed this shill account by participating in some tasks before the experiment.
4. Low Price, No Shill: same as treatment 1 except that the reward is 100 yuan.
5. Low Price, Shill without credit: same as treatment 2 except that the reward is 100 yuan.
6. Low Price, Shill with credit: same as treatment 3 except that the reward is 100 yuan.

Experiment Procedure

We posted 148 tasks on Taskcn from June 3 - June 22, 2009, 8 tasks per day. We select a single winner for each task. To avoid reputation effects from the askers' side, we used different Taskcn identities for each task by creating

[1] See http://www-personal.umich.edu/~liuxiao/pdf/taskcnfield20110525_EC_updated.pdf

148 new accounts. Therefore, each task was posted by a unique user ID with no history. After a task was posted, any user could participate and submit their answers within 7 days. After the seventh day, we selected one answer as the winner and the shill is never selected as the winner.

HYPOTHESES

In this section, we describe our hypotheses comparing users' behaviors between different treatments based on our theoretical predictions. We are interested in two outcome measures: participation and answer quality. Specifically, we have the following hypotheses:

1. A task with a high reward attracts more submissions than a task with a low reward.
2. A task with a high reward attracts answers of higher quality than a task with a low reward.
3. The early entry of a high-quality answer (the shill answer) will deter the entry of others, consequently, the total amount of participation in shill treatments will be less than in the no-shill treatments. Furthermore, this shill effect on the number of participation is more salient in shill-with-credit treatments than in shill-without-credit treatments.
4. The average answer quality will be lower in shill treatments than no-shill treatments, especially in shill-with-credit treatments.

RESULTS

Rating Procedure

The rating protocol here is similar to the one used in Chen et al. (2010). For the translation tasks, nine Chinese graduate students were recruited from the University of Michigan. The majority of them were masters students at the School of Information. As the school requires a TOEFL score of at least 600 when admitting international students, they all had relatively high reading and writing skills in English as non-native speakers. For the programming tasks, three Chinese graduate students were recruited from the University of Michigan, School of Information. All of them had an undergraduate major in computer science and several years of experience in web programming.

As there were 3671 translation answers in total, the nine raters were randomly assigned to three different rating groups. Raters within each group independently rated the same question-answer pairs. On the other hand, the three programming raters rated all programming tasks due to the small number of answers for programming tasks.

All raters followed the same rating procedure for each question-answer pair. For each question, we randomly selected one machine translation from the answer pool as well as all non-machine translations. All the rating questions can be found on the corresponding author's website.[2]

To improve the reliability of students' ratings, we conducted training sessions before the rating sessions began. For translation tasks, we gave raters one sample personal statement and company introduction, then asked them to rate the difficulty of both questions. We also gave them two answers for each question and asked them to rate each answer's quality. One of the answers was written by the personal statement provider or our two undergraduate research assistants, and the other was randomly drawn from the answers that we received from the pilot session. For the programming task, we follow the same procedure with two sample questions. In addition, to help raters develop and refine their own personal rating scales instead of encouraging consensus among them, we asked them to individually give reasons for their rating scores for each question-answer pair.

From October 2009 to February 2010, we conducted 45 rating sessions at the University of Michigan, School of Information Laboratory. Each session lasted two hours to prevent fatigue. Students were paid a flat fee of $15 per hour to compensate them for their time. We used intra-class correlation coefficients to measure inter-rater reliability. The intraclass correlation coefficient (ICC) for translation answer quality in each group is: 0.90, 0.88, 0.68 and it is 0.49 for programming answers, representing a good-to-excellent reliability.

The Price Effect

We first examine how different prices affect participation. Due to the existence of machine translations and answers copied from others, we examined the quantity of all submissions and the quantity of human answers separately (for translation tasks). The criteria for human answers, which represent a certain amount of effort, are therefore: (1) not machine-translated; and (2) not copied from others.

We find that no matter which type of the answer they receive, translation tasks in high-price treatments always have more submissions compared to tasks in low-price treatments (Average Number of Submissions/Question: All Answers: 35 vs. 26, $p<0.01$; Human Answers: 6 vs. 3, $p<0.01$, one-sided two-sample t-test).

Next, we examine the price effect on human answer quality. Using an ordered probit specification with standard error clustered at the question level, we find that the average quality of human translation answers is higher in high-price treatments than in low-price treatments (Average Median Quality: 5.06 vs. 4.76, $p=0.028$, one-sided). Consistently, the quality of the best translation answer is higher in high-

[2] See http://www-personal.umich.edu/~liuxiao/pdf/ taskcnfield20110525_EC_updated.pdf

price treatments than in low-price treatments (Average Median Quality: 6.04 vs. 5.67, p=0.012, one-sided).

Regarding programming tasks, we find consistent price effect on both participation and answer quality (Submission: p=0.068; Quality: p=0.027, one-sided).

The Reserve-Price (Shill) Effect

In this section, we analyze the reserve-price (shill) effect on answerers' behaviors. Due to the significant higher answer quality of shill answers than others, we expect to observe fewer submissions and lower answer quality in shill treatments compared to no-shill treatments as it leaves little room for others to improve, particularly if the shill answer is posted by a user with credits.

Although there is no significant shill effect on participation, the average quality of human translation answers is lower in the two shill treatments than in the no-shill treatments (Average Median Quality: 4.71 (4.80) vs. 5.29, p<0.01, one-sided). Furthermore, the quality of the best answer is lower in the shill treatments compared to the no-shill treatments (Average Median Quality: 5.69 (5.63) vs. 6.24, p<0.01, one-sided).

Individuals' Entry

Lastly, we investigate how different types of users choose to participate in tasks. As the mode of users' credits, which approximate their winning experiences, is 0, we categorize users into two types: (1) experienced users whose credits are greater than 0; (2) inexperienced users who never win before the experiment.

We compute the proportion of high-price tasks undertaken by each user and find that experienced users are less likely to choose high price tasks than inexperienced users (All Answers: 0.70 vs. 0.75, p=0.02; Human Answers: 0.75 vs. 0.90, p<0.01, one-sided two-sample t-test). In addition, using Kolmogorov-Smirno Test, we find the cumulative distribution function (cdf) of average user credit for each question between the high price treatment and the low price treatment is significantly different (p=0.031, one-sided).

Next, we examine how the existence of shill answers influences each type of user's behaviors. Specifically, experienced users are significantly less likely to choose tasks with shill answers than inexperienced users (All Answers: 0.77 vs. 0.81, p=0.017; Human Answers: 0.75 vs. 0.88, p<0.01, one-sided two-sample t-test). Moreover, the cdf of average user credit for each question between shill and no shill treatments is significant ($p=0.047$, one-sided). These two results indicate that experienced users are more strategic when choosing tasks with different monetary incentives and more likely to observe others' behaviors before participation.

CONCLUSION

In this paper, we studied different design features of crowdsourcing sites with an all-pay auction mechanism by conducting a field experiment on Taskcn. By manipulating the monetary incentive and the existence of a reserve price in the form of a good shill answer, we find that higher price induces more participation and higher answer quality. Furthermore, the existence of a reserve price lowers the answer quality in general and the individual analysis shows that it is because of the less entry from experienced users. In addition, to increase their winning probability, experienced users are more strategic regarding their participation in tasks with different monetary incentives (resp. reserve prices). We hope our results can have implications in the design of crowdsourcing sites.

REFERENCES

Chen, Y., Ho, T., & Kim, Y. (In Press). Knowledge Market Design: A Field Experiment on Google Answers. *Journal of Public Economics Theory.*

Harper, F. M. , Raban, D., Rafaeli, S., & Konstan, J. A. (2008). Predictors of Answer Quality in Online Q&A Sites. *CHI2008.*

Howe, J. (2006). The Rise of Crowdscourcing, *Wired.*

Mason, W., & Watts, D.J. (2009). Financial Incentives and the "Performance of Crowds". *KDD-HCOMP.*

Nov, O., Naaman, M., & Ye, C. (2008). What Drives Content Tagging: the Case of Photos on Flickr. *Proceedings of the 26th annual SIGCHI conference on Human factors in computing systems.*

Kleeman, F., Voss, G.G., & Rieder, K. (2008). Un(der)paid Innovators: The Commercial Utilization of Consumer Work through Crowdsourcing. *Science, Technology & Innovation Studies, 4 (1), 5–26.*

Segev, E., & Sela, A. (2011). Sequential All-pay Auctions with Head Starts. *Working Paper.*

Yang, J., Adamic, L.A., & Ackerman, M.S. (2008). Crowdsourcing and Knowledge Sharing: Strategic User Behavior on Taskcn. *Proceedings of the 8th ACM conference on Electronic Commerce.*

Zhang, J., Adamic, L.A., Ackerman, M. S., & Bakshy, E. (2008). Everyone knows something: Examining knowledge sharing on Yahoo Answers. *World Wide Web(WWW'08).*

.