Winning by Learning?

An Exploratory Study of Knowledge Sharing and Usage on Crowdsourcing Platforms

Sulin Ba, Yuan Jin, Brian Lee, and Jan Stallaert

University of Connecticut

- Preliminary Draft –

Please do not circulate

Abstract

A crowdsourcing platform enables solution seekers to use contributions from a group of online users, usually by organizing crowdsourcing contests through which online users contribute to the solution generation process. Knowledge sharing on a crowdsourcing platform could play an important role in the process of crowdsourcing contests and contestants generating high-quality solutions. On one hand, more knowledge resources could lower the participation cost and help improve crowdsourcing outcomes. On the other hand, the shared knowledge may also interrupt contestants' independent solution search processes and limit contestants' creativity. Specifically, knowledge sharing can affect crowdsourcing outcomes via the parallel path effect and stimulating effect in crowdsourcing contests. This study provides an in-depth analysis on how on a crowdsourcing platform the knowledge sharing process impacts contestants' innovation process and contest outcome quality. We first examine how different dimensions of the shared knowledge affect crowdsourcing contests. We find that high knowledge sharing originality negatively impacts the outcome by decreasing the parallel path effect and stimulating effect, while high knowledge sharing quality positively impacts the outcome by increasing the stimulating effect. Subsequent analyses using a finite mixture model indicate that contestants vary in their ways of using knowledge and hence are influenced by knowledge sharing differently.

1. Introduction

In recent years, crowdsourcing platforms have become increasingly popular among businesses that seek new ideas for their product and technology development. A tournament-based crowdsourcing platform, such as Kaggle and TopCoder, holds crowdsourcing contests to organize a group of online users who provide solutions to the solution seeker. Online users participate in the crowdsourcing contest and search for solutions independently, and solution seekers select the most satisfactory submissions and offer monetary awards to the winners. Contestants in a crowdsourcing contest compete with each other to win the prize, but at the same time the knowledge sharing process on the platform provides a way of community cooperation. For example, people can have discussions and share information in the online forums on these crowdsourcing platforms. In this paper, we investigate how knowledge sharing on the crowdsourcing platforms impacts crowdsourcing contest outcomes and contestant performance.

Knowledge sharing could play an important role in the process of crowdsourcing contests whose basic aim is to look for high-quality ideas and solutions. Knowledge sharing refers to the provision of information and know-how to help or collaborate with others to develop ideas and solve problems (Pulakos et al. 2003, Cummings 2004). The knowledge management literature has shown that knowledge sharing can facilitate new product development and improve firm performance (e.g. Hansen 2002, Arthur and Huntley 2005, Collins and Smith 2006).

Knowledge sharing on crowdsourcing platforms can occur via forum discussions, information exchange among contestants, questions and answers between the seekers and the contestants, etc. However, for crowdsourcing platforms, the supposed beneficial effect of knowledge sharing might be more nuanced. On one hand, as the knowledge sharing process facilitates information exchange among contestants and provides knowledge resources to lower contestants' participation costs, it can help improve contestant performance and crowdsourcing outcomes. On the other hand, knowledge sharing may also be a burden to both knowledge sharers and recipients if the knowledge is not transferred efficiently (Hendriks 1999), thus distracting contestants from fully focusing on solution development and hurting the crowdsourcing outcome and performance instead.

For instance, Kaggle.com, a predictive modeling crowdsourcing platform on which seekers post their data and solvers write code to produce prediction models, features a type of discussion boards called Kernels where people can share coding scripts associated with a contest. The shared coding scripts are for various purposes such as data cleansing, data visualization and algorithm development. On Kaggle Kernels, the coding languages include Julia, Python, R and SQLite. Figure 1 shows the screenshot of a data visualization script shared for Expedia Hotel Recommendations contest[1]. As the code conveys knowledge directly, sharing coding scripts represents a typical form of knowledge sharing. People may learn from the shared scripts and thus improve their performance, while the other possibility is that the shared scripts distract them and limit their creativity.

[1] This contest was held between April 15, 2016 and June 10, 2016. Expedia provided a random selection of logs of customer behavior and challenged contestants to contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups.

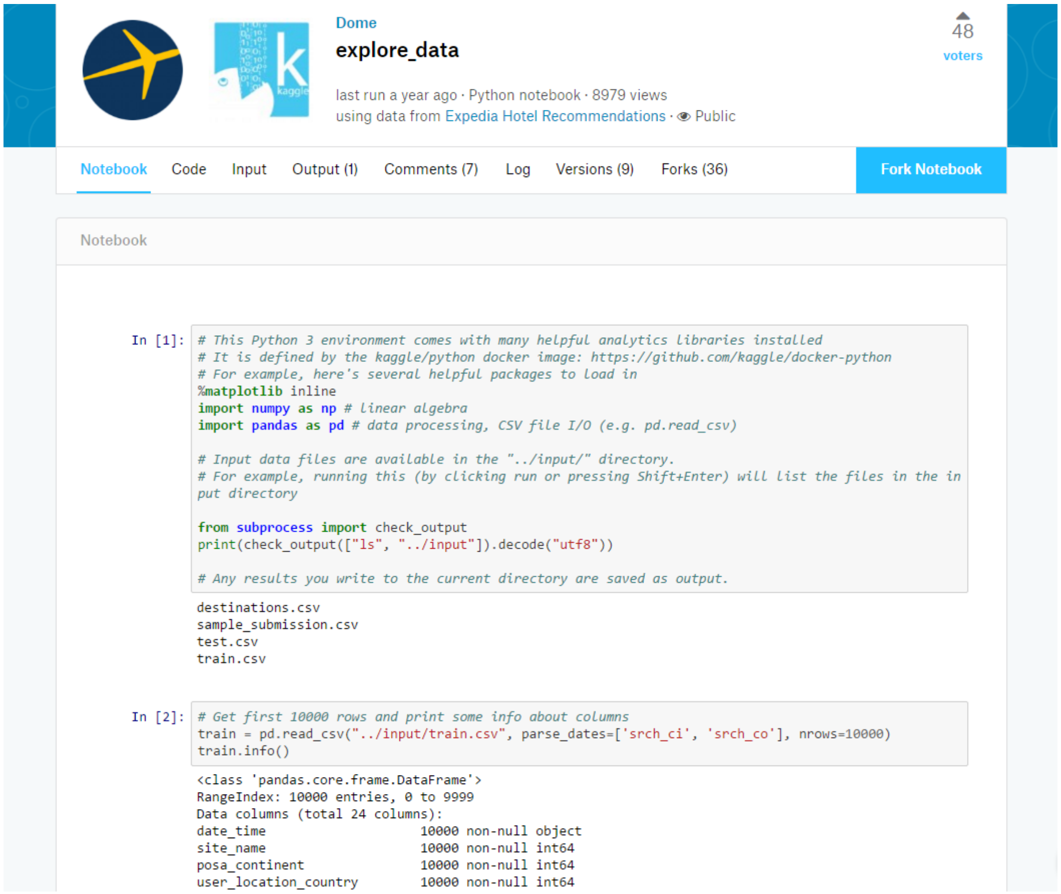

Figure 1. An Example of Kaggle Scripts

To dig into how exactly the shared knowledge affects crowdsourcing contests, we need to look at different dimensions of knowledge sharing, as they may have different effects on crowdsourcing outcome and performance. In this study we focus on two dimensions: originality and quality of knowledge sharing. The originality of knowledge sharing refers to how much original knowledge has been shared on the platform, measuring the volume of the new knowledge added. Originality is one of the most important characteristics of knowledge (Ismail Al-Alawi et al. 2007), but in an online setting where users have easy access to the knowledge shared by others, people sometimes choose to develop the already existing knowledge and share the derivative knowledge. Compared with the derivative knowledge, original knowledge shared contributes more to the knowledge sharing process as it adds more to the body of the knowledge with completely new ideas and opportunities for future development. So instead of simply including the absolute quantity of the shared knowledge, we use the originality of the shared knowledge to show the volume of overall knowledge contribution. Quality is another critical dimension of knowledge sharing, reflecting the value and usefulness of the knowledge. On a crowdsourcing platform, knowledge sharing can be beneficial or detrimental to the crowdsourcing contests, and the two dimensions (originality and quality) may affect the crowdsourcing outcome and performance in different ways.

Besides the direct impact, knowledge sharing could also influence crowdsourcing contests and contestants indirectly by moderating the parallel path effect and stimulating effect. Prior research has identified the parallel path effect and stimulating effect as key contributing factors for the success of crowdsourcing. The parallel path effect states that more contestants providing independent searches for solutions increases the likelihood of the contest getting a desirable result (Boudreau et al. 2011). The stimulating effect describes a situation where a larger number of participants (i.e., a larger crowd) improves outcomes, as larger tournaments induce higher levels of competition and consequently more effort input from contestants, who want to increase their chance of winning by exerting more effort (Boudreau et al. 2016). Knowledge sharing could moderate the parallel path effect and stimulating effect in crowdsourcing contests. As knowledge sharing may interrupt contestants' independent solution search processes by affecting individuals' idea generation with shared views and methods, it could possibly limit contestants' creativity and hence lower the variability of contestants' idea generation processes. Such lowered variability could decrease the parallel path effect, creating a less extreme value outcome (Terwiesch and Xu 2008). The shared knowledge may also influence contestants' perception of the contest competition so that it can moderate the stimulating effect as well. Understanding the possible interaction between the two effects and knowledge sharing is important for crowdsourcing contests, as knowledge sharing might alter the working mechanisms of crowdsourcing.

In this study, we analyze the impact of knowledge sharing on crowdsourcing contests and contestants in depth. Although knowledge sharing exists on many platforms, the influence and potential contribution of sharing knowledge on crowdsourcing platforms are novel to the literature. Specifically, we address these questions: 1) How do different dimensions (originality and quality) of knowledge sharing influence contestants' innovation progress and the overall crowdsourcing outcome? 2) How does knowledge sharing moderate the parallel path effect and stimulating effect? 3) Does knowledge sharing affect contestants' crowdsourcing performance differently given the contestant heterogeneity?

2. Related Literature and Theoretical Background

This research is mainly related to two streams of literature: crowdsourcing platforms and contests, and knowledge sharing and management. Previous crowdsourcing research has examined what factors determine the success of crowdsourcing contests and the quality of the generated solutions. These factors include contest characteristics such as contestants' participation (Terwiesch and Xu 2008, Boudreau et al. 2011, Boudreau et al. 2016) and reward size (Huang et al. 2014), and contestant characteristics such as experience and expertise (Jeppesen and Lakhani 2010, Bayus 2012) and team structures (Dissanayake et al. 2015). However, there is no study that specifically investigates the role of knowledge sharing on a crowdsourcing platform in influencing the contest outcome and contestant performance.

Knowledge sharing is an activity through which people provide information to others in the process of developing ideas or solving problems (Pulakos et al. 2003, Cummings 2004). It can occur via communications with other people, or knowledge documenting and capturing for others (Pulakos et al. 2003, Cummings 2004). The value of knowledge grows when the knowledge is shared: when one shares knowledge with other units, a linear growth occurs as those units gain information; at the same time an exponential growth occurs as the units feed back questions, amplifications and modifications that add further value for the original sender (Quinn et al. 1996, Cabrera and Cabrera 2002).

As such, members of an organization can contribute to knowledge application and amplification, which would add to the organization's competitive advantage, by sharing knowledge (Jackson et al. 2006). Prior research has shown that knowledge sharing helps organizations exploit knowledge-based resources (Cabrera and Cabrera 2005), and is positively related to both organizational performance (such as sales growth, profitability of a new product, and completion speed of product development) and individual performance (e.g., Hansen 2002, Cummings 2004, Arthur and Huntley 2005, Collins and Smith 2006).

However, the impact of knowledge sharing may not always be beneficial. Hendriks (1999) argues that the shared knowledge can be augmented only when people truly learn from each other, and knowledge sharing may prove detrimental if inadequate representations of knowledge are transferred among people. The negative effect of knowledge sharing might be more obvious on crowdsourcing platforms, as within a limited time, contestants are supposed to spend as much effort as possible on working on their own solutions, but participating in and contributing to knowledge sharing could divert the effort. However, even with the possible negative effect, knowledge sharing might still influence crowdsourcing contests in a positive way, since it can facilitate information exchange among contestants and provide knowledge resources to lower contestants' participation costs.

Different dimensions of knowledge sharing may have different effects on crowdsourcing outcomes and contestant performance. Wasko and Faraj (2005) study the motivation behind people voluntarily contributing knowledge and helping others through electronic networks. In their work, the knowledge contribution to an online community is measured by the volume of contribution and the helpfulness of the knowledge. Similarly, in this paper we include the originality and quality of the shared knowledge as the focal dimensions. Originality shows the volume of new knowledge shared on the platform. Quality reflects the value and usefulness of the shared knowledge on the platform.

2.1 Moderating Role of Knowledge Sharing

Knowledge sharing could moderate the parallel path effect and stimulating effect, which are two important factors contributing to the success of crowdsourcing contests. The parallel path effect states that when there are more submitted solutions (i.e., more independent solution searches) from contestants, the chance of the contest getting a desirable result will increase (Boudreau et al. 2011). If we view each submission from contestants as a random "draw" from an underlying distribution of possible quality of outcomes, more submissions would lead to a greater chance of uncovering an extreme value outcome (Boudreau et al. 2011). The contest seeks submissions of the highest quality, so with a higher chance of getting an extreme value outcome, its final outcome would likely have a higher quality.

The stimulating effect describes a situation where an increase in the number of participants (i.e., crowd size) improves outcomes, as the high competition level would stimulate contestants, especially the high-ability ones, to exert more effort so as to increase their winning probability (Boudreau et al. 2016). Boudreau et al. (2016) find that contestants' average performance declines as the competition - measured by the number of competitors - increases in a contest. However, their result also shows that the competition has different effects on contestants of different types. For the contestants of low ability, whether the competition is intense or not does not really affect their performance, but the competition would affect the performance of medium-ability contestants negatively. This is because for these contestants whose ability is not sufficiently high, a more intense competition would signal a lower chance for them to win, resulting in these contestants exerting less effort which leads to worse performance. But for the contestants with sufficiently high ability, the competition would do the opposite: the negative effect would then decline to even become positive for the contestants of very high ability, because these contestants would obtain more motivation to pay effort when they face stronger competition so as to increase their chance of winning.

As indicated by the parallel path effect, if there are more independent innovation attempts, the likelihood of getting a high-quality outcome would be higher (Boudreau et al. 2011). That is, the higher the submission variability, the stronger the parallel path effect. Knowledge sharing, as it may interrupt the independent solution search process and lower the variability, could reduce the parallel path effect. Yet there is another possibility: the shared knowledge would save contestants' initial set-up time and enhance their solution searching efficiency, thus lead to more diverse submissions as contestants now have more time to fully develop their ideas. This could help increase the parallel path effect. The stimulating effect is measured by the crowd size in a contest and mainly works on high-ability contestants (List et al. 2014, Boudreau et al. 2016). Knowledge sharing could amplify this effect, because more shared knowledge (especially high-quality shared knowledge) observed by a contestant would make the contestant perceive a higher possibility of his/her competitors getting improved. Combined with the competition reflected by the crowd size, this perceived possibility can motivate contestants to exert even more effort to outrun other contestants who benefit from knowledge sharing. But at the same time, knowledge sharing may bring distractions to contestants. Even though contestants increase effort input, they would place some of the increased effort on trying to absorb the shared knowledge (which may not work or fit the contestant's own way of solving the problem after all) instead of working on their own solutions, leading to an attenuated stimulating effect. Also, if the competition indicated by the combination of knowledge sharing and large crowd size appears to be overwhelming for even high-ability contestants, the stimulating effect would be reduced with contestants losing incentives to exert more effort.

2.2 Contest-level Model

Both high originality and high quality of knowledge sharing on crowdsourcing platforms could affect crowdsourcing contest outcomes, and we would like to examine whether they have direct influence on the contest outcome and if they do, whether the impact is positive or negative.

As discussed in § 2.1, both the parallel path effect and stimulating effect could be moderated by knowledge sharing. So we also examine how exactly knowledge sharing moderates the two effects and hence indirectly affects crowdsourcing outcomes.

Figure 2 summarizes the contest-level model.

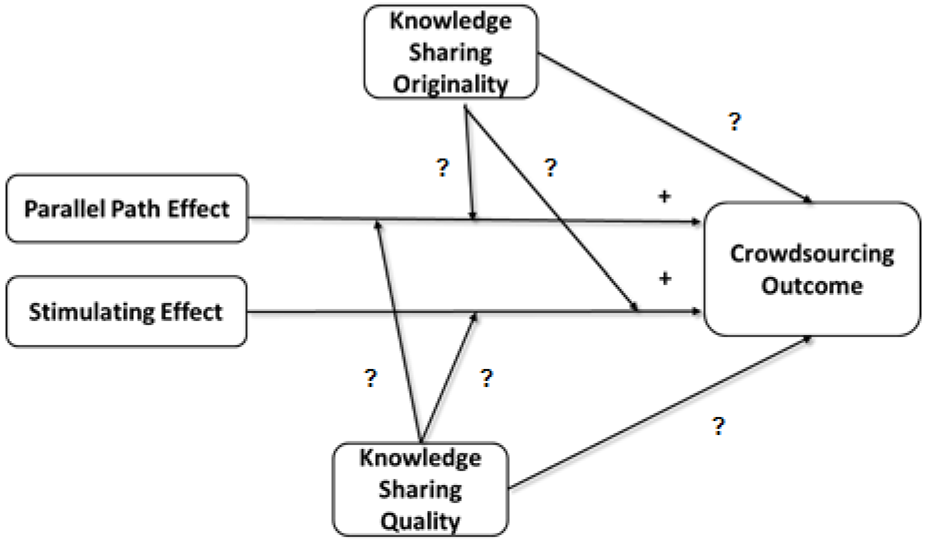

Figure 2. Contest-level Model

2.3 Contestant-level Model

Although knowledge sharing may influence the contest outcome by affecting contestants' behavior and performance, the impact of knowledge sharing on contestants is not necessarily identical to that on contests. The existence of parallel path effect may amplify the effect of knowledge sharing on the variability of submitted solutions, leading to a possible inconsistency between the effect on contestant performance and the effect on contest outcomes. At the same time, as contestants may use or learn from the shared knowledge in different ways, knowledge sharing could also have different influence on different types of contestants. With different types of contestants performing in and contributing to the contest differently, the way of knowledge sharing influencing a contestant's performance can be different from it affecting the overall contest outcome.

In addition to the contest-level model, we also examine the correlation between knowledge sharing and contestants' crowdsourcing performance with a contestant-level model, in which the two dimensions of knowledge sharing could influence a contestant's performance and moderate the stimulating effect[2].

Figure 3 summarizes the contestant-level model.

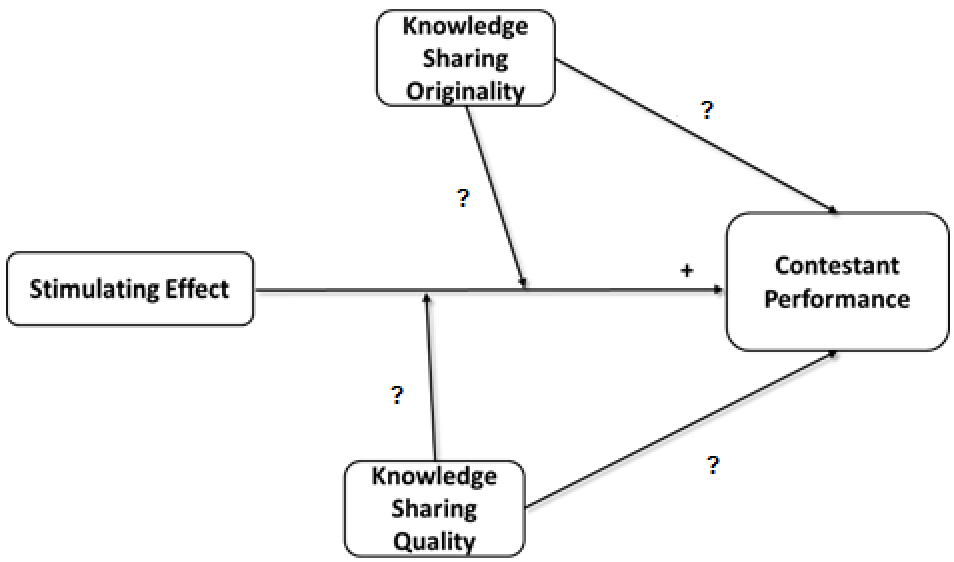

Figure 3. Contestant-level Model

3. Research Context

In this study, we carry out our analysis using an archival data set from Kaggle. Founded in 2010, Kaggle is a platform for predictive modeling and analytics crowdsourcing contests. In Kaggle contests, companies and researchers post their data, and statisticians and data miners from all over the world compete to produce the best models. As of May 2016, Kaggle had over 536,000

[2] Parallel path effect is associated with the variability of all contestants' submissions, so it does not work on individual contestants.

registered users from 194 countries, forming the largest and most diverse data community in the world. Kaggle contests regularly attract over a thousand teams and individuals. To find the best models, the solution seeker first prepares the data and a description of the problem, and the platform will then release the problem as a contest to public. Kaggle members can participate in the contest and become contestants. Contestants experiment with different techniques and compete against each other to produce the best models. They can submit multiple entries and get scored for each entry before a contest ends. At the end of the contest, they select one or multiple submissions for the final evaluation. Among these final submissions, one or several are chosen to be the ultimate contest winner(s). For most competitions, submitted entries are scored immediately based on their predictive accuracy and summarized on a live leaderboard. Kaggle also enables a feature called Kernels, which allows users to write, run, and publicly share their code on the platform.

Our sample consists of 25 contests that were held between May 15, 2014 and April 15, 2016 and enabled script sharing feature. Table 1 shows the summary statistics for the contests in our sample.

Table 1. Contest Summary Statistics (N=25)

	Mean	Std. Dev	Min	Median	Max
No. of Contestants	1,557.76	1,191.73	316	1,256	5,022
No. of Valid Submitted Solutions	26,769.48	23,613.73	1,049	23,239	93,558
Award Size (in dollar)	25,860	21,801.31	250	25,000	100,000
Contest Length (in day)	69.6	18.88	38	66	126
No. of Scripts	758.04	577.40	5	668	2,351

4. Contest-level Analysis

Using Kaggle data, we examine how knowledge sharing influences the crowdsourcing contest outcome directly and indirectly via the parallel path effect and stimulating effect.

To measure the originality and quality of knowledge sharing, we use the Kernels information associated with the contests as indicators. Kernels on Kaggle is a special discussion board where Kaggle members can post scripts and methods related to each contest. Members can also vote for and comment on a script, and the number of votes for a script can reflect the quality of this script. In the data set, we also have information on whether a script cites another script[3]. If a script has not cited any other script, we call it an original script. We use script originality instead of script quantity (i.e., how many scripts, including both the original and derivative ones, have been shared) because on Kaggle Kernels usually a derivative script makes only minor changes to the original script, and hence it is the original scripts that really add to the body of the shared knowledge. If a script scores in the top 5% in terms of the adjusted number of votes[4] it has received in total, we call that script a high-vote script. We use the ratio of the number of original scripts to the number of all scripts ($O_{j,t-1}$) created on or before contest day $t-1$ to show the originality of the shared knowledge for contest j on day t, and the ratio of the number of high-vote scripts to the number of all scripts ($Q_{j,t-1}$) created on or before contest day $t-1$ to show the quality of the

[3] On Kaggle, using and citing an existing script can be done by a function called "fork". The data show whether a script forks (cites) another script and if it does, which script it forks.

[4] We divide the absolute number of votes by the number of views the script has received to adjust the number of votes.

shared knowledge for contest *j* on day *t*.

We include the total number of teams that have participated in the contest on or before contest day $t-1$ ($TTeam_{j,t-1}$) to show the overall crowd size, which reflects the extent of the stimulating effect (List et al. 2014, Boudreau et al. 2016). We use the submission volume to measure the parallel path effect ($STeam_{jt}$). Instead of the number of teams that submit entries on contest day *t,* we use the ratio of this number to the total number of teams that have participated in the contest on or before contest day $t-1$, to get rid of the possible correlation between the number of submitting teams and the total number of participating teams (if there are more participating teams, the number of submitting teams would probably be larger too).

To understand the effects of knowledge sharing on the outcome of the contests, in the contest-level analysis we use the maximum standardized private score ($Score_{jt}^{max}$) as the dependent variable to measure the best outcome quality from a contest on a specific day. Kaggle provides two sets of scores as feedback to contestants' submissions. When submissions are made during the course of a contest, contestants receive public scores as a feedback; at the end of the contest, contestants select one or multiple entries as their final submissions, and receive a private score for each submission as the final evaluation. Compared with the public score, the private score is a more reliable indicator of the submission quality as it is usually calculated based on a larger test set and is used for final evaluation and winner decision. We standardize the private scores to make submissions across contests comparable. Maximum standardized private score is the maximum among all entries for contest *j* that are submitted on day *t*.

Furthermore, we also want to examine how knowledge sharing influences the average solution quality and the submission variability. So we use the average standardized private score ($Score_{jt}^{avg}$) on contest day *t* and the standard deviation of standardized private scores ($Score_{jt}^{std}$) on contest day *t* as the dependent variables in the other two specifications.

To control the natural learning process of the crowd over time, we include the progress variable P_{jt} of contest *j* on contest day *t* ($P_{jt} = \frac{t}{Total\ No.of\ Contest\ Days}$). We also include the total number of scripts ($S_{j,t-1}$) created on or before contest day $t-1$ as a control variable. Fixed effects for contest *j* (η_j) are included to control for the heterogeneity of the contests. ε_{jt} is the idiosyncratic error for contest *j* and day *t*.

Except for the progress variable P_{jt}, all other variables are log transformed because they are skewed. And all knowledge sharing variables ($O_{j,t-1}$, $Q_{j,t-1}$, and $S_{j,t-1}$), the submission volume ($STeam_{jt}$) and the crowd size ($TTeam_{j,t-1}$) variables are centered on their means. The contest-level model examines the correlation between the crowdsourcing outcome from a contest and the knowledge sharing variables.

$$\begin{aligned} Score_{jt} = {} & \beta_1 O_{j,t-1} + \beta_2 Q_{j,t-1} + \beta_3 STeam_{jt} + \beta_4 TTeam_{j,t-1} + \beta_5 O_{j,t-1} \times STeam_{jt} + \beta_6 O_{j,t-1} \\ & \times TTeam_{j,t-1} + \beta_7 Q_{j,t-1} \times STeam_{jt} + \beta_8 Q_{j,t-1} \times TTeam_{j,t-1} + \beta_9 P_{jt} \\ & + \beta_{10} S_{j,t-1} + \eta_j + \varepsilon_{jt} \end{aligned}$$

In the contest-level analysis, we have three specifications to examine the effect of knowledge sharing on the contest best outcome, the average solution quality and the solution variability respectively. The results are provided in Table 2. The model in Column (1), as our main model, shows how knowledge sharing affects the daily maximum score and moderates the parallel path effect and stimulating effect. Column (2) and (3) further check the effect of knowledge sharing on average solution quality and submission variability.

Table 2. Contest-level Results

Dependent Variable ($Score_{jt}$)	Max. Score ($Score_{jt}^{max}$)	Avg. Score ($Score_{jt}^{avg}$)	Std Dev. Score ($Score_{jt}^{std}$)
Shared Knowledge Originality	-0.001	-0.021	0.494**
($O_{j,t-1}$)	(0.026)	(0.037)	(0.176)
Shared Knowledge Quality	-0.010	-0.003	-0.525*
($Q_{j,t-1}$)	(0.026)	(0.030)	(0.294)
Submission Volume	0.043*	0.026	0.318**
($STeam_{jt}$)	(0.022)	(0.019)	(0.141)
Crowd Size	0.010**	0.014**	0.045**
($TTeam_{j,t-1}$)	(0.004)	(0.005)	(0.022)
$O_{j,t-1} \times STeam_{jt}$	-0.218***	-0.076	-2.797***
	(0.076)	(0.151)	(0.882)
$O_{j,t-1} \times TTeam_{j,t-1}$	-0.029**	-0.022	-0.271***
	(0.013)	(0.018)	(0.083)
$Q_{j,t-1} \times STeam_{jt}$	0.236	0.089	4.092**
	(0.139)	(0.180)	(1.488)
$Q_{j,t-1} \times TTeam_{j,t-1}$	0.042**	0.005	0.462***
	(0.020)	(0.021)	(0.137)
Learning Progress	0.003	-0.002	0.025
(P_{jt})	(0.008)	(0.012)	(0.043)
No. of All Scripts	0.002	0.007**	-0.031**
($S_{j,t-1}$)	(0.002)	(0.003)	(0.014)
# Observations	1,691	1,691	1,691

# Contests	25	25	25
R^2	0.792	0.411	0.191

*** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value <0.1

As indicated by Column (1), the main effects of both knowledge sharing dimensions are not significant, but both dimensions show significant moderating effect on the parallel path effect or stimulating effect. High originality of the shared knowledge will decrease both the parallel path effect and stimulating effect, and high quality of the shared knowledge will increase the stimulating effect. This indicates that a large volume of knowledge sharing will negatively affect outcomes by reducing the parallel path effect and stimulating effect. One possible explanation is that with a large knowledge sharing volume, there is a higher chance that a contestant essentially changes or even abandons his/her own idea according to a shared script, leading to an interrupted solution search process and a decreased parallel path effect. Also, a large volume of knowledge sharing consumes a lot of contestants' effort that should have been otherwise spent on solution search. At the same time, it shows contestants' active and committed participation in the contest, which, combined with a large crowd size, signals such an intense competition that dilutes contestants' incentives to improve their performance, leading to a weakened stimulating effect. High quality of the shared knowledge, on the other hand, amplifies the stimulating effect, indicating the high-quality knowledge only works positively on the crowdsourcing contest when the crowd is sufficiently large and diverse.

The net effect of knowledge sharing is mainly determined by the indirect effect via the parallel path effect and stimulating effect. We find that the high originality of knowledge sharing has a negative impact while the high quality of knowledge sharing has a positive impact, especially when the submission volume or the crowd size is large. The implication here is that the platform should try to improve the quality of the shared knowledge while at the same time avoid information overload. When the participation of a contest increases, the importance of ensuring high quality and avoiding information overload also increases.

Column (2) shows how knowledge sharing influences average outcome quality from a contest. The results show that the knowledge sharing process does not really affect the overall quality of the submitted solutions. Column (3) shows how knowledge sharing works on the variability of submitted solutions. The two dimensions of knowledge sharing affect the submission variability in a very similar way as they affect the extreme value outcome, though the main effect of knowledge sharing originality is positive for submission variability, which means high originality of knowledge sharing can improve the submission variability when the contest participation is low. The results in Column 3 provide some insight into how knowledge sharing influences the crowdsourcing outcome: it can influence the crowdsourcing outcome via its effect on the submission variability, which is a major determinant for the success of crowdsourcing contests.

5. Contestant-level Analysis

Besides the effect of knowledge sharing on the overall contest outcome, we are also interested in how knowledge sharing would affect individual contestants' performance. In this section, we provide a contestant-level analysis based on the contestant-level model. We include 1,464 contestants from the 25 contests who have at least voted once for the scripts shared for the contest that they participate in, which shows the evidence that these contestants have read the shared scripts.

In this contestant-level analysis, instead of the maximum score achieved by all contestants in a contest, we use the highest standardized private score that contestant *i* gets on day *t* ($Score_{ijt}$) to measure the contestant's performance. In the contest-level analysis, we measure the two dimensions of knowledge sharing by using the ratios of original/high-vote scripts to all scripts shared for a contest by a specific day. Here to measure the pure influence of the knowledge shared by others on contestant i, we adjust these two knowledge sharing variables by deducting the original/high-vote scripts shared by contestant *i* him/herself on or before contest day $t-1$ from the numerators and deducting all the scripts shared by contestant *i* him/herself on or before contest day $t-1$ from the denominators. We adjust the number of all scripts ($S_{ij,t-1}$) in the same manner. Also, now instead of the contest progress variable P_{jt}, we include the contestant progress variable P_{ijt}, measured by the ratio of the ordinal number of submission day *t* to the number of all submission days of contestant *i* in contest j.

In the contestant-level analysis, we first run a pooled regression which estimates the average effect of knowledge sharing for all contestants in the sample. However, an assumption for the pooled regression is that all contestants are homogeneous, which may not be true since there could exist heterogeneity in contestants' patterns of using the shared knowledge and hence how they are influenced by knowledge sharing. So we further employ the finite mixture model to capture unobserved individual heterogeneity and to segment contestants. A finite mixture model assumes that contestants can be heterogeneous. It can use contestants' observed crowdsourcing performance together with the knowledge sharing covariates to infer their unobserved knowledge-using types. Within each contestant segment produced by the finite mixture model, the contestants are assumed to be homogeneous.

Due to the difficulty of estimating fixed effects in the finite mixture model, instead of including the fixed effects for contestant i in contest j (η_{ij}), we normalize each variable (including all dependent and independent ones) by deducting its mean for contestant i in contest j from the original variable to control for the heterogeneity of the contestants and contests. It is worth mentioning that we also checked the fixed effect model for the pooled contestant-level analysis, and the results of including fixed effects are essentially the same with the results of using normalized variables.

$$Score_{ijt} = \gamma_1 O_{ij,t-1} + \gamma_2 Q_{ij,t-1} + \gamma_3 TTeam_{j,t-1} + \gamma_4 O_{ij,t-1} \times TTeam_{j,t-1} + \gamma_5 Q_{ij,t-1} \times TTeam_{j,t-1} + \gamma_6 P_{ijt} + \gamma_7 S_{ij,t-1} + \varepsilon_{ijt}$$

In the contestant-level analysis, we examine the correlation between knowledge sharing and the score a contestant receives on a day. Table 3 shows the results of the pooled regression together with the finite mixture model results. From the results of the pooled analysis, we find overall the high quality of knowledge sharing has a positive impact on a contestant's crowdsourcing performance and amplifies the stimulating effect on a contestant, while the high originality of knowledge sharing decreases the stimulating effect.

Applying the finite mixture model, we identify three types of contestants that exhibit different knowledge-using tendencies[5]. The 3-segment finite mixture model results show us a different picture from what we find in the pooled analysis. Knowledge sharing has differential impact on the contestants in different segments. Segment 1 consists of 432 contestants (29.51%) out of the 1,464 contestants in our sample. The main effect of high knowledge sharing originality is positive on the contestants in this segment but it decreases the stimulating effect, and high knowledge sharing quality has a negative effect. The magnitudes of all the effects are quite small. Segment 2 consists of 447 contestants (30.53%). Similar to Segment 1, knowledge sharing originality has a positive direct effect but decreases the stimulating effect. But for the contestants in this segment, high knowledge sharing quality has a positive indirect effect via the stimulating effect. Segment 3 consists of 585 contestants (39.36%). High knowledge sharing originality hurts the performance of the contestants in this segment by reducing the stimulating effect. But high knowledge sharing quality works positively on them, improving their performance and increasing the stimulating effect for them.

We compare the contestants' characteristics in different segments to understand what

[5] A critical issue of applying the finite mixture model is to decide the optimal number of segments. In this analysis, we compare the 2-segment, 3-segment, 4-segment and 5-segment finite mixture models based on likelihood-based information criteria including AIC and BIC to determine the optimal number of segments. Although 4-segment and 5-segment models can still slightly improve AIC and BIC, the improvement is comparatively small and the major estimation difference between segments is in the coefficients of control variables, not in those of our focal variables. So we choose the 3-segment model considering the model's interpretability and parsimony.

contestant type each segment represents. Table 4 shows the means and standard deviations (shown in parentheses) of contestants' best scores, rankings[6] and submission totals for each segment. Contestants in Segment 1 has the lowest average best score as well as the lowest ranking. They also make submissions least frequently. Contestants in Segment 2 has the highest average best score and the highest ranking. They submit most frequently. Contestants in Segment 3 show in-between performance. Generally speaking, contestants in Segment 1, Segment 2 and Segment 3 provide low performance, high performance and medium performance respectively.

Table 3. Contestant-level Results

Dependent Variable ($Score_{ijt}$)	Overall	Segment 1	Segment 2	Segment 3
Shared Knowledge Originality	-0.008*	2.19×10^{-4} **	0.017***	-0.022
($O_{ij,t-1}$)	(0.004)	(8.94×10^{-5})	(0.001)	(0.036)
Shared Knowledge Quality	0.022***	-3.82×10^{-4} ***	-0.001	0.089***
($Q_{ij,t-1}$)	(0.007)	(1.45×10^{-4})	(0.001)	(0.020)
Crowd Size	0.004***	2.33×10^{-4} ***	0.004***	0.006***
($TTeam_{j,t-1}$)	(0.001)	(1.59×10^{-5})	(0.000)	(0.002)
$O_{ij,t-1} \times TTeam_{j,t-1}$	-0.012***	-4.25×10^{-4} ***	-0.010***	-0.045***
	(0.002)	(4.61×10^{-5})	(0.000)	(0.011)
$Q_{ij,t-1} \times TTeam_{j,t-1}$	0.011***	6.59×10^{-5}	0.004***	0.066***
	(0.004)	(7.74×10^{-5})	(0.001)	(0.001)
Learning Progress	0.013***	2.97×10^{-4} ***	0.004***	0.020***
(P_{ijt})	(0.001)	(2.84×10^{-5})	(0.000)	(0.001)
No. of All Scripts	-0.001***	-2.69×10^{-7}	-0.001***	-0.003***
($S_{ij,t-1}$)	(0.000)	(7.16×10^{-6})	(0.000)	(0.001)
# Observations	13,462	2,591	5,441	5,430
# Contestants	1,464	432	447	585
Segment Share		29.51%	30.53%	39.96%

*** indicates p-value<0.01, ** indicates p-value<0.05, * indicates p-value <0.1.

[6] We use the ranking percentile for comparison. The ranking percentile is calculated by dividing a contestant's ranking in a contest by the number of all ranked contestants in this contest.

Table 4. Contestant Segment Comparison[7]

	Avg. Best Score	Avg. Number of Submissions	Avg. Ranking
Segment 1	0.019 (0.692)	6.234 (9.121)	0.476 (0.285)
Segment 2	0.401 (0.326)	12.353 (10.674)	0.287 (0.225)
Segment 3	0.370 (0.468)	9.516 (7.829)	0.403 (0.239)

Combining the performance statistics with the finite mixture model results, we find that for most of the contestants, high originality of knowledge sharing could have a negative impact. It indeed improves high-performance contestants' performance directly, but it reduces the stimulating effect for all three contestant types. High quality of knowledge sharing, on the other hand, positively influences both the medium-performance and high-performance contestants. It mainly improves medium-performance contestants' performance, but negatively impacts low-performance contestants. Depending on the platform's goal, the platform can choose appropriate strategy to adjust the two knowledge sharing dimensions to help certain types of contestants improve performance.

6. Discussion and Conclusion

In this study, we first examine how different dimensions of knowledge sharing influence the crowdsourcing contests and how they moderate the parallel path effect and stimulating effect.

---

[7] Except that the difference between the average best scores of Segment 2 and Segment 3 is not significant at the 0.1 level (P-value=0.118), all other statistics in the table have significant between-segment differences at the 0.01 level.

We find that high originality of knowledge decreases the parallel path effect and stimulating effect, while high quality of knowledge sharing increases the stimulating effect for a contest.

Then we take one step further to examine how individual contestants are affected by knowledge sharing. From the pooled analysis, we find that contestant performance is positively influenced by the high quality of knowledge sharing. But the pooled contestant-level analysis only shows the overall impact of knowledge sharing on contestant performance; it is of importance to differentiate between different types of contestants in using knowledge and explore how these contestant types contribute differently to the contest final outcome. So we apply a finite mixture model to identify and compare contestants' knowledge-using types, and find knowledge sharing indeed has different effects on contestants, and the different contestant types provide different performance in the contest.

The contestants of high performance would benefit from a large volume of original shared knowledge, as they are able to absorb a wide variety of shared ideas and further obtain inspiration from the various knowledge. On the contrary, medium-performance contestants would be distracted by the large volume of new knowledge and hence be negatively impacted by the high knowledge sharing originality, especially when the competition is intense. High quality of knowledge sharing would amplify the stimulating effect in most cases, and benefit the medium-performance contestants mostly. But it hurts the performance of those low-performance contestants. These results provide managerial implications for a crowdsourcing platform to understand how knowledge sharing influences contestants of different types and to apply appropriate strategies for implementing and encouraging knowledge sharing based on its goals.

In this study, we provide an in-depth analysis on the impact of knowledge sharing on crowdsourcing contests. The main contributions of this research include: 1) adding to the crowdsourcing literature by providing a detailed analysis on the effect of knowledge sharing on crowdsourcing contest outcomes and contestants' crowdsourcing performance; 2) providing an examination of how knowledge sharing alters crowdsourcing mechanisms by moderating the parallel path effect and stimulating effect; 3) adding to the knowledge management literature by demonstrating that both positive and negative effects of knowledge sharing exist on crowdsourcing platforms, which is different from other scenarios where knowledge sharing is usually shown to be beneficial; and 4) providing managerial implications for crowdsourcing platforms to identify contestants' knowledge-using types and understand how shared knowledge impacts different types.

References

Arthur, J. B., & Huntley, C. L. 2005. "Ramping up the organizational learning curve: Assessing the impact of deliberate learning on organizational performance under gainsharing," *Academy of Management Journal* (48:6), pp. 1159-1170.

Bayus, B. L. 2013. "Crowdsourcing New Product Ideas over Time: An Analysis of the Dell IdeaStorm Community," *Management Science* (59:1), pp. 226–244.

Boudreau, K. J., Lacetera, N., and Lakhani, K. R. 2011. "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis," *Management Science* (57:5), pp. 843–863.

Boudreau, K. J., Helfat, C. E., Lakhani, K. R., and Menietti, M. 2016. “Performance Responses to Competition across Skill-levels in Rank Order Tournaments: Field Evidence and Implications for Tournament Design.,” *RAND Journal of Economics* (47:1), pp. 140–165.

Cabrera, A., & Cabrera, E. F. 2002. “Knowledge-sharing dilemmas,” *Organization studies* (23:5), pp. 687-710.

Cabrera, E. F., & Cabrera, A. 2005. “Fostering knowledge sharing through people management practices,” *The international journal of human resource management* (16:5), pp. 720-735

Collins, C. J., & Smith, K. G. 2006. “Knowledge exchange and combination: The role of human resource practices in the performance of high-technology firms,” *Academy of management journal* (49:3), pp. 544-560.

Cummings, J. N. 2004. “Work groups, structural diversity, and knowledge sharing in a global organization,” *Management science*, (50:3), pp. 352-364.

Dissanayake, I., Zhang, J. I. E., and Gu, B. I. N. 2015. “Task Division for Team Success in Crowdsourcing Contests: Resource Allocation and Alignment Effects,” *Journal of Management Information Systems* (32:2), pp. 8–39.

Hansen, M. T. 2002. “Knowledge networks: Explaining effective knowledge sharing in multiunit companies,” *Organization science* (13:3), pp. 232-248.

Hendriks, P. 1999. “Why share knowledge? The influence of ICT on the motivation for knowledge sharing,” *Knowledge and process management* (6:2), pp. 91.

Huang, Y., Vir Singh, P., & Srinivasan, K. 2014. “Crowdsourcing new product ideas under consumer learning,” *Management science* (60:9), pp. 2138-2159.

Ismail Al-Alawi, A., Yousif Al-Marzooqi, N., & Fraidoon Mohammed, Y. 2007. "Organizational culture and knowledge sharing: critical success factors," *Journal of knowledge management* (11:2), pp. 22-42.

Jackson, S. E., Chuang, C. H., Harden, E. E., Jiang, Y., & Joseph, J. M. 2006. "Toward developing human resource management systems for knowledge-intensive teamwork," *Research in personnel and human resources management*, (25:6), pp. 27-70.

Jeppesen, L. B., and Lakhani, K. R. 2010. "Marginality and Problem Solving Effectiveness in Broadcast Search," *Organization Science* (21:5), pp. 1016–1033.

List, J., Van Soest, D., Stoop, J., & Zhou, H. 2014. *On the role of group size in tournaments: Theory and evidence from lab and field experiments* (No. w20008). National Bureau of Economic Research.

Pulakos, E.D., & Dorsey, D.W., and Borman, W.C., 2003. "Hiring for knowledge-based competition," in *Managing knowledge for sustained competitive advantage: Designing strategies for effective human resource management*, S. E. Jackson, A. DeNisi, & M. A. Hitt (eds.), John Wiley & Sons, pp. 155-177.

Quinn, J. B., Anderson, P., & Finkelstein, S. 1996. "Leveraging intellect," *The Academy of Management Executive* (10:3), pp. 7-27.

Terwiesch, C., and Xu, Y. 2008. "Innovation contests, open innovation, and multiagent problem solving," *Management Science* (54:9), pp. 1529–1543.

Wang, S., & Noe, R. A. 2010. "Knowledge sharing: A review and directions for future research," *Human Resource Management Review* (20:2), pp. 115-131.

Wasko, M. M., & Faraj, S. 2005. "Why should I share? Examining social capital and knowledge contribution in electronic networks of practice," *MIS quarterly*, pp. 35-57.