Does Exposure to Shared Solutions Lead to Better Outcomes? An Empirical Investigation in Online Crowdsourcing Contests

Jingbo Hou
Arizona State University
jhou27@asu.edu

Pei-yu Chen
Arizona State University
peiyu.chen@asu.edu

Bin Gu
Boston University
bgu@bu.edu

Abstract

Crowdsourcing contests provide an effective way to elicit novel ideas and creative solutions from collective intelligence. A key design feature of crowdsourcing contests is the competition between contest participants to complete a specific task with financial awards to the winner(s). In recent years, some crowdsourcing contest platforms provide options to contest participants for solution sharing during the competition. This study intends to evaluate the influence of exposure to shared solutions on different stakeholders, including the team, and the requester. Our study employs a multiple-level panel data from a large online crowdsourcing platform, Kaggle.com, to examine these effects. For teams, exposure to shared solutions helps new entrant teams to jump-start and help teams to achieve better performance in the subsequent submissions, and the teams' skill level negatively moderates these positive effects. For requesters, allowing solution sharing has both benefits and costs in terms of improving the best performance of the crowd. We highlight the theoretical implications of the study and provide practical suggestions for crowdsourcing contest platforms to help them decide whether to allow solution sharing during the competition.

1. Introduction

Crowdsourcing contests have become more and more attractive for organizations to generate ideas and solve problems because of the unprecedented scale and diversified background of the labor pool they provide [1]. An increasing number of organizations, including governments (e.g., Health and Human Services Department), research institutes (e.g., NASA), large enterprises (e.g., General Electric, LG), have started to employ crowdsourcing contests to enable their research and development process[1].

Online crowdsourcing contest platforms facilitate access to a large labor pool and provide an easily accessible and efficient way for companies to obtain ideas and/or solutions [2]. To attract more participants and achieve better crowdsourcing outcomes, crowdsourcing contest platforms explore ways in an attempt to lower entry barriers and reduce participation costs for contestants. For example, kaggle.com, the largest crowdsourcing platform focusing on data science-related problems, provides a mechanism that allows participants to share their intermediary solutions during the contest. To motivate contestants to share high-quality solutions, Kaggle awards the authors of popular shared solutions. During the competition, the contestants take shared solutions as a benchmark or inspiration to aid their innovations [2]. To come up with solutions, contestants first search over the solution space in the exploration stage. For example, when contestants first join a competition, they explore the task requirement specified by the requesters and explore all the existing solutions either provided by other solvers or provided outside the platform. Then in the exploitation stage, they exploit the most promising area found in the exploration stage. Allowing solution sharing may cause contestants to shift their effort from exploring new directions independently to exploiting the shared solutions.

From the exploitation perspective, using shared solutions helps contestants gain the required skill and knowledge quickly and boost their performance [3]. Meanwhile, contestants can allocate more time to the critical components of the solution because they do not need to duplicate effort on reinventing the basic components. As a result, the final solution could be improved. In a way, that is how society has progressed by building upon shared knowledge.

From the exploration perspective, allowing solution sharing may also have some unintended effects. For example, allowing solution sharing may disincentivize

[1] https://www.ideaconnection.com/open-innovation-success/

URI: https://hdl.handle.net/10125/71408
978-0-9981331-4-0

contestants from exploring new directions independently. One of the primary goals of high-skilled contestants for innovation is to win the award/prize with the minimum effort. To economize their effort level, contestants may shift the effort from exploring new directions to exploiting the existing shared solutions. Shared solutions may cause contestants to think inside the boundary set by existing solutions, which is detrimental in the innovation process. The above discussions suggest that there are tradeoffs in allowing solution sharing for different stakeholders. The consequence of solution sharing is still unclear.

This study intends to evaluate the influence of solution sharing in crowdsourcing contests on different stakeholders, including the participating teams and the requester, which initiates the contest. Specifically, we address the following research questions:

1. How does exposure to shared solutions influence the performance of participating teams at different skill levels?

2. How does exposure to shared solutions influence participating teams' parallel path effect and the contest outcomes?

We find that solution sharing is overall beneficial for the crowdsourcing platform, the requester, and the teams. For participating teams, solution sharing helps new entrant teams to jump-start and helps existing teams to achieve better performance during the contests. Low-skilled teams benefit more from solution sharing functionality. For requesters, allowing solution sharing has both benefits and costs in terms of improving the best performance of the crowd. The findings have important implications for crowdsourcing contest platforms.

2. Hypotheses Development

2.1 Impact of Solution Sharing on Teams' First-Submission Performance

When new entrant teams, which are inexperienced in a contest, are exposed to the shared solutions, they can learn from these shared solutions to gain the required skills and domain knowledge [4]-[5]. The learning behavior is similar to exploitation in solution search literature [6]-[7]. Exploitation here means that contestants can exploit the promising intermediary solutions shared by others. Exploiting existing knowledge helps individuals to get workable solutions [8], have more innovative [9], and more effective solutions [10], and achieve more secure performance outcomes [11]. When the number of shared solutions is bigger, new entrant teams to a contest can learn more to boost their first-submission performance.

It is worth noting that the positive effect of shared solutions on teams' first-submission performance might be heterogeneous for teams at different skill levels. For high-skilled teams, they may benefit less from learning from the shared solutions. Because they already mastered the required skills and domain knowledge to start the contest independently, and/or shared solutions may hinder these teams from thinking beyond the boundary set by these shared solutions, which is called 'adverse fixation effect' in the innovation literature (e.g., [12]-[14]). Therefore, the high-skilled teams will benefit less from shared solutions to jump-start their first-submission performance.

H1a. The number of shared solutions in a contest improves the teams' first-submission performance (i.e., the performance of the new entrant teams to a contest). H1b. The lower the skill level of new entrant teams, the greater the effect of shared solutions on the teams' first-submission performance.

2.2. Impact of Solution Sharing on Teams' Subsequent Performance

Teams can broaden their skill sets and improve their solutions during the contest through observational learning from the shared solutions. Teams who have low performance in the contest have much space to improve by learning from shared solutions. However, high-performance teams in a contest are less likely to keep improving their performance through learning because they already have outstanding performance compared with their peers. At the same time, the fixation effects influence high-skilled teams more because they have the required skills to come up with high-quality solutions if they think independently. When the high-skilled teams learn from the solutions proposed by others, they may get stuck by the shared solutions and/or they may incorporate some inappropriate (even detrimental) features in their own solutions unintentionally [15]. Therefore exposure to shared solutions may benefit high-skilled teams less compared with average-skilled teams during the competition.

H2a. The number of solutions used by teams has a positive effect on their subsequent solution performance. H2b. The lower the teams' historical performance in a contest, the greater the positive effect of solution usage on their subsequent solution performance.

2.3. Impact of Solution Sharing on the Best-performance of the Crowd

Parallel path effects predict when teams develop solutions independently and parallelly, the increased number of teams leads to a higher chance that the contest might get an exceptionally high-quality solution [16] – [21]. In our study, the data science task (e.g., finding dark matter in the universe) is highly complicated and uncertain, and this high uncertainty amplifies the parallel path effects [21]. So adding more teams increases the chance that the requester gets an exceptionally high-quality solution.

During the contest, exposure to solutions shared by others helps high-skilled teams to come up with revolutionary creative solutions. First, when teams work on high-quality shared solutions as benchmarks, they can save time from duplicating the basic components of the same task and use these saved time to work on the critical components of the task. Second, when teams can observe multiple shared solutions, they have a decent chance of finding superior solutions that provide them with perspectives from a new angle. Jeppesen and Lakhani [22] empirically found that the provision of winning solutions is positively related to the distance between the solver's expertise and the focal field of the problem. The rationale behind this phenomenon is that when the current direction of the solution does not work, having a perspective outside the current field domain may help to generate an effective solution. From this perspective, exposure to the existing superior solutions shared by others with different skillsets may help teams to have alternative perspectives. Thirdly, teams can compare their solutions with the existing shared solutions and then reflect and revise their own solutions. This reflection process is essential in experiential learning.

However, shared solutions may hinder the independent revolutionary thinking of creative teams and make them conform to shared solutions. Best-performance solutions, which beat all other solutions, are likely to be revolutionary creative solutions. When exposed to the shared solutions, the high-skilled teams who have the potential to come up with these best-performance solutions might be more vulnerable to the fixation effects. For these highly creative teams, being exposed to solutions shared by others may trigger conformity effects and hinder them from proposing extremely creative solutions [13]. After being exposed to shared solutions, these shared solutions are involuntarily retrieved in mind and cannot be deliberately rejected [13]. To sum up, the shared solutions may attenuate the parallel path effects because the shared solutions may discourage the most creative teams from thinking independently.

H3a. The number of teams increases the best solution performance of all teams (i.e., the parallel path effects). H3b. The number of shared solutions increases the best solution performance of all teams. H3c. The more the shared solutions, the smaller the effect of the number of teams on the best solution performance.

3. Research Context and Data

Our study employs a multiple-level panel data from a large online crowdsourcing platform, Kaggle.com, to examine the effects of solution sharing on different stakeholders. Kaggle is a crowdsourcing contest platform, which allows requesters to post contests and seek solutions for their data science tasks. There are multiple reasons why we chose Kaggle as our research context. First, Kaggle provides 'kernel' functionality to encourage contestants to share solutions during the contest. 'Kaggle kernel' means shared scripts/IPython Notebooks/R Markdown, combining the programming environment, input, code, and output. Contestants can share the intermediary solutions by making their kernels public. The shared kernels are available to all teams. Kaggle tries to use the kernel function to help contestants to manage and share their data science work. Second, Kaggle provides a well-organized and daily updated archival dataset (i.e., Meta Kaggle), enabling scholars to examine the effects of exposure to shared solutions at multiple levels empirically.

Until Nov 2019 (the time we got our data), Kaggle has held 360 different public contests since its launch. In our study, we only include 237 contests that provide a monetary reward. We exclude the contests held for new contestants in Kaggle for educational purposes, consistent with previous studies [23]. In the Kaggle contests, solutions are submitted based on teams, and the majority teams are single-member teams. Each contest gives the upper bound of the number of team members.

4. Variables, Model Specification and Results

Our analysis is at different levels, including contest level and team level. To quantify the impact of solution sharing on different stakeholders, we aggregate the team-level panel data to contest level to test our H1 and H3, and we use the single-member-team-level analysis to test our H2. We employ the fixed effects as our main identification strategies for all analyses, including the team-specific and contest-specific fixed effects. These fixed effects help us account for average differences across teams and contests in any observable or unobservable predictors, such as the team's job experience and contest complexity.

4.1. Impact of Solution Sharing on Team's First-Submission Performance

To examine how the shared solutions help new entrant teams to get better performance, we only consider the first submission of each team, and our analysis is at the contest-day level. More specifically, we aggregate the team-level first submission performance to the contest level to study how the number of shared solutions available before teams submit their first solution influences these teams' first-submission performance. During the time window of the analysis, each contest attracts 13 teams on average per day (the Std. Dev. is 22.2). In our analysis, we aggregate the performance of all teams' first-submission solutions of a contest on the same day by using the daily average performance scores. To ensure that new entrant teams' daily performance scores are comparable across different contests, we calculate the normalized performance scores. Specifically, we use Equation 1 to calculate the normalized average daily performance scores of teams' first-submission solutions. The normalized performance scores are independent of the contest.

$$AvgScore_{jt} = \frac{AvgOriginalscore_{jt} - mean(AvgOriginalscore_{j,t=1\ldots T})}{\sigma_{AvgOriginalScore_{j,t=1\ldots T}}} \times 1(j) \quad (1)$$

, where j indexes contest j, t denotes time t, and $1(\cdot)$ is the indicator function, and the definition of this indicator function is:

$$1(j) = \begin{cases} 1 & if \text{ performance } measurement\ is\ accuracy\ rate\ based \\ -1 & if \text{ performance } measurement\ is\ error\ rate\ based \end{cases}$$

There are different types of evaluation metrics, including the accuracy type and error rate type. In the accuracy type of metrics, the high-performance score represents high performance. However, in the error rate type of metrics, the smaller score represents better performance. Here we use an indicator function to ensure that the high normalized performance score represents high performance.

We use the performance in the private leaderboard to measure the team's first-submission performance. Kaggle calculates the public score by a relatively small portion of the holdout set (e.g., 10%) and calculates the private leaderboard by a more substantial portion of the holdout set (e.g., 90%). We use the performance in the public leaderboard as the robustness check, and the results are consistent. The variable of interest is the number of solutions available at time t-1 (i.e., $\mathrm{KernelNum}_{j,t-1}$). We also control for the number of posts in the forum ($\mathrm{ForumPostNum}_{j,t-1}$) at t-1. Note that using variables $\mathrm{KernelNum}_{j,t-1}$ and $\mathrm{ForumPostNum}_{j,t-1}$ effectively address the potential concern of reverse causality.

Teams may show strategic behavior during the competition. For example, teams exert more effort toward the end of the competition to avoid submission wars [24]. High-quality teams will submit their solutions later than inexperienced ones, and high-quality teams are less likely to enter tasks when a high-quality solution has already been submitted [25]. Failure to account for these timing strategies may bias our results. For instance, if experienced team leaders strategically wait until the end to submit their solutions (e.g., [25]), these teams are more likely to have better performance no matter they can learn from the shared solutions or not.

To account for the timing strategy mentioned above, we control for the time elapsed ($\mathrm{TimeElapsed}_{j,t}$) of a contest and the average quality of team leaders in our analysis. The time elapsed is measured by the percentage of contest time elapsed since the start divided by the contest's total duration [24]. The measurement of the average quality of team leaders of new entrant teams includes the $NewUserRate_{jt}$ and $AvgRanking_{jt}$.

Further, we control the contest-specific fixed effects to address systematic contest differences that are invariant over time. Failure to control for these contest-specific characteristics may lead to the spurious correlation between exposure to shared solutions and team performance. For example, suppose some attractive contests (e.g., contests with the higher budget) draw more high-skilled entrants and shared solutions simultaneously. In that case, the regression results will be biased without controlling the contest-specific fixed effects.

We specify our model as Equation 2 to test our H1 and present the definition and the summary statistics of the variables in Panel A of Table 1. We log-transformed some of our independent variables due to the high data skewness [33].

Table 1 Variables Used to Test H1 and H3 (Contest-level Analysis)

Panel A: Variables Used to Test H1 (Contest-Day Level)					
Dependent Variable	N	mean	sd	min	max

$AvgPublicScore\_NE_{jt}$	The normalized average score of new entrant teams' first submissions in contest j on day t in the public leaderboard (*only used in the robustness check*)	12630	0.000	0.992	-12.04	8.460
$AvgPrivateScore\_NE_{jt}$	The normalized average score of new entrant teams' first submissions in contest j on day t in the private leaderboard	12450	0.000	0.993	-12.04	8.481
Variable of Interest						
$KernelNum_{jt-1}$	The number of shared solutions in contest j on or before t − 1	15753	107.6	244.0	0	2611
Control Variables						
$ForumPostNum_{jt-1}$	The number of available posts in the forum of contest j on or before t − 1	15753	68.91	86.71	0	648
$TimeElapsed_{jt}$	The percentage of contest (j) time elapsed as of current day t	15753	0.508	0.289	0.00143	1
$NewUserRate_{jt}$	The percent of new entrant teams' leaders without contest experience in contest j on day t	13178	0.548	0.297	0	1
$AvgRanking_{jt}$	The average historical contest rank percentile of the leaders of new entrant teams' in contest j on day t	10860	0.429	0.145	0.00152	1
Panel B: Variables Used to Test H3 (Contest-Week Level)						
Dependent Variable						
$MaxPublicScore_{jt}$	Normalized best performance score of all teams in contest j at or before week t in the public leaderboard (*only used in the robustness check*)	845	0.000	0.697	-2.474	2.153
$MaxPrivateScore_{jt}$	Normalized best performance score of all teams in contest j at or before week t in the private leaderboard	860	0.000	.695	-2.341	2.171
Variable of Interest						
$TeamNum_{jt}$	The number of teams in contest j at or before week t	860	1087.6	1176.6	16	8491
$KernelNum_{jt}$	The number of shared kernels in contest j at or before week t	860	247.6	339.1	0	2528

$$DailyScore\_NE_{jt} = \beta \times \ln(KernelNum_{jt-1} + 1) + \beta \times \ln(ForumPostNum_{jt-1} + 1) + \beta \times TimeElapsed_{jt} + \beta \times TeamLeaderQuality\_NE_{jt} + \alpha_j + \epsilon_{jt} \quad (2)$$

Table 2 the Impact of Solution Sharing on New Entrants' Performance

Model	(1)
DV	$AvgPrivateScore\_NE_{jt}$
$\ln(KernelNum_{jt-1} + 1)$	0.0968*
	(0.0509)
$\ln(ForumPostNum_{jt-1} + 1)$	0.124**
	(0.0620)
$\ln(NewUserRate_{jt} + 1)$	-0.531***
	(0.0871)
$\ln(AvgRanking_{jt} + 1)$	-1.158***
	(0.124)
$TimeElapsed_{jt}$	0.0604
	(0.114)
Constant	-0.166
	(0.162)
Observations	10,687
R-squared	0.040
Number of Contests	212
Contest FE	YES

*Notes. Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1;*

As shown in Table 2, the lagged number of shared solutions is positively related to the teams' first-submission performance in the private leaderboard, supporting our H1a. Our results indicate when the number of shared solutions increases by 10%, the new-entrant teams' normalized performance would increase by 1%. For requesters in the crowdsourcing contest, they care more about the daily best-performance. To this end, we also examine the effect of shared solutions on the daily best performance across all first-submission solutions in the robustness check section. Similarly, the number of shared solutions is positively related to the daily best performance across all first-submission solutions.

After examining the effect of exposure to shared solutions on the team's first-submission performance, we are still interested in whether this effect is different across the contestants with different qualities. We leverage the regression quantile method proposed by Machado and Santos Silva [26] to examine this heterogeneous effect. Whereas the OLS estimates the conditional mean of the new entrant's performance across predictor variables, the regression quantiles estimate the conditional quantiles of the response variable. The higher quantiles represent the performance of a higher skill level. We run the regression quantiles for the 10th, 30th, 50th, 70th, and the 90th percentiles of the distribution and find the positive effect of shared solutions is stronger on low-performance distribution than on high-performance distribution, which supports our H1b.

Table 3 the Impact of Solution Sharing on New Entrants' Average Performance (Regression Quantile)

Model	(1)	(2)	(3)	(4)	(5)
Outcome Variable (quantile):	$AvgPrivateScore\_NE_{jt}$				
	q10	q30	q50	q70	q90
$\ln(KernelNum_{jt-1}+1)$	0.118**	0.105***	0.0963***	0.0891***	0.0797**
	(0.0564)	(0.0339)	(0.0235)	(0.0225)	(0.0330)
$\ln(ForumPostNum_{jt-1}+1)$	0.194**	0.148***	0.117***	0.0911***	0.0573
	(0.0883)	(0.0530)	(0.0368)	(0.0353)	(0.0517)
$\ln(NewUserRate_{jt}+1)$	-0.258	-0.433***	-0.547***	-0.645***	-0.773***
	(0.181)	(0.109)	(0.0754)	(0.0723)	(0.106)
log_AvgRankPercentile_Sub	-1.389***	-1.234***	-1.132***	-1.045***	-0.932***
	(0.249)	(0.149)	(0.104)	(0.0995)	(0.146)
$TimeElapsed_{jt}$	-0.229	-0.0450	0.0754	0.179**	0.313***
	(0.182)	(0.109)	(0.0760)	(0.0729)	(0.107)
Observations	10,104	10,104	10,104	10,104	10,104
Contest FE	YES	YES	YES	YES	YES

*Notes. Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1*

4.2. Impact of Solution Sharing on Teams' Subsequent Performance

Even though contestants can observe all the shared solutions, they may not choose to use all the solutions given their limited cognitive ability. Therefore, we operationalize 'exposure to the shared solutions' as 'the number of votes given to shared solutions.' Giving votes to solutions that contestants think is useful is a social norm in Kaggle, and Kaggle encourages all contestants to follow this norm. The voted solutions are more likely to be exploited by contestants.

The solution voting behavior happens at the contestant level, but the performance is measured at the team level. Leveraging the single-member teams can help us avoid the level mismatch issue. Therefore, we conduct a single-member-team week level analysis. In Kaggle competition, most teams are single-member teams, as indicated by the average team size 1.028. Focusing on the single-member teams does not hinder the generalizability of our results. As a robustness check of our contestant-level exposure measurement, we operationalize 'exposure to the shared solutions' using 'the number of shared solutions,' and the results are consistent.

We organize our data at a weekly level because most of the contestants update their submission entries on a weekly base. In our time window of analysis, each contestant submits 1.2 entries per week on average. The dependent variable is the normalized average performance of the single-member team i in contest j at week t. We use Equation 3 to calculate the normalized average daily performance scores of each single-member team (i.e., a contestant at a contest) at a different time:

$$AvgScore_{ijt} = \frac{AvgOriginalscore_{ijt} - mean(AvgOriginalscore_{i=1\ldots I,j,t=1\ldots T})}{\sigma_{AvgOriginalScore_{i=1\ldots I,j,t=1\ldots T}}} \times 1(j) \quad (3)$$

We use Equation 4 as the model specification to test our H2. In the analysis, we control for the team fixed effects to account for the team-specific factors, including the team's capability. We also account for the teams' effort level through controlling for the lagged number of submissions and the lagged performance in the public leaderboard during the contest [33]. We present the definition and the summary statistics of the variables in Table 4.

As shown in Table 5, our results indicate a positive relationship between exposure to shared solutions and the team's performance. The lagged performance negatively moderates this positive relationship, which supports our H2a and H2b. Exposure to one more solution increases the team's standardized performance by 0.6%.

Table 4 Variables Used to Test H2 (Single-Member-Team-level Analysis)

Dependent Variable		N	mean	sd	min	max
$AvgPublicScore_{ijt}$	The normalized average score of single-member team i in contest j at week t in the public leaderboard (*only used in the robustness check*).	2539178	-0.146	1.173	-104.6	39.46
$AvgPrivateScore_{ijt}$	The normalized average score of single-member team i in contest j at week t in the private leaderboard.	2539178	-0.133	1.188	-109.8	12.01

$SubmissionNum_{ijt}$	The number of code submissions by single-member team i in contest j at week t.	2539178	1.198	23.28	0	11563
Variable of Interest						
$VoteNum_{ijt-1}$	The number of votes given by single-member team i to the shared solutions in contest j at week $t-1$.	2539178	0.0411	0.402	0	113
Control Variables						
$CommentNum_{ijt-1}$	The number of comments given by single-member team i to the shared solutions in contest j at week $t-1$.	2539178	0.00805	0.141	0	31
$\mathrm{SubmissionNum}_{ij(t-1)}$	The number of solution submissions by single-member team i in contest j at week $t-1$.					
$TimeElapsed_{ijt}$	The percentage of contest (j) time elapsed as of current week t for single-member team i	2539178	0.542	0.289	0.00990	1

$$AvgScore_{ijt} = \beta_1 \times VoteNum_{ijt-1} + \beta_2 \times SubmissionNum_{ij(t-1)} + \beta_3 \times TimeElapse_{ijt} + \beta_4 \times \ AvgPublicScore_{ijt-1} + \alpha_{ij} + \ \epsilon_{ijt} \quad (4)$$

Table 5 the Impact of Solution Sharing on Teams' Performance

Model	(1)	(2)
DV	$AvgPrivateScore$	$AvgPrivateScore_{ijt}$
$VoteNum_{ijt-1}$	0.00643***	0.0167***
	(0.00113)	(0.00376)
$VoteNum_{ijt-1} \times AvgPublicScore_{ijt-1}$		-0.0720***
		(0.00933)
$SubmissionNum_{ij(t-1)}$	6.97e-05	0.671***
	(9.30e-05)	(0.0165)
$AvgPublicScore_{ijt-1}$	0.669***	7.69e-05
	(0.0165)	(9.37e-05)
$TimeElapsed_{ijt}$	0.0269***	0.0282***
	(0.00285)	(0.00283)
Constant		-0.0510***
		(0.00339)
Observations	2,326,971	2,326,971
R-squared	0.424	0.425
Number of Teams	208,197	208,197
Team FE	YES	YES

*Notes. Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1;*

4.3. Solution Sharing Has Competing Effects on Best Solutions

We employ the panel VAR (e.g., [27]) approach to account for the bilateral effect between the performance of the best solutions and the team number, while controlling for contest-specific heterogeneity. Panel VAR technique combines the traditional VAR approach, which treats all variables as endogenous variables, and the panel-data approach, which can control panel-specific heterogeneity. Therefore, Panel VAR models allow us to account for the bilateral effect and control the unobserved contest-specific heterogeneity. We specify our models as follows:

$$Y_{jt} = \ \Gamma Y_{jt-1} + \alpha_j + e_{jt}$$

Where Y_{jt} is a four-variable vector $\{\ln(TeamNum_{jt}+1), \ln(KernelNum_{jt}+1), \ln(TeamNum_{jt}+1) \times \ln(KernelNum_{jt}+1), MaxPrivateScore_{jt}\ (or\ MaxPublicScore_{ij}\ )\}$. The crowd's best performance measurement is the highest score of all teams until a given week. We organize our contest panel data for the panel VAR analysis at the contest-week level.

Kaggle has two different contest formats, including the simple competition and two-stage competition format. As for two-stage competitions, Meta Kaggle does not provide teams' performance on the first stage. Therefore, we only include 91 simple competitions launched after Kaggle introduced kernel functionality in our contest-week level panel VAR analysis. We present the definition and the summary statistics of the variables in Panel B of Table 1. The variables of interest include the number of teams, the number of shared solutions, and the interaction term between the number of teams and the number of shared solutions. TeamNum and KernelNum variables are log-transformed due to the skewness of the data.

For our panel VAR analysis, we first need to decide the period of lags in our model. We use the model selection criteria to help us find out the optimal period of lags. We calculate the model selectin criteria measures for first to fifth-order panel VAR using the first six lags of $\{\ln(TeamNum_{jt}+1), \ln(KernelNum_{jt}+1),\ \ln(TeamNum_{jt}+1) \times \ln(KernelNum_{jt}+1), MaxPrivateScore_{jt}\ (or\ MaxPublicScore_{ij}\ )\}$. Based on the criteria proposed by Andrews and Lu [28], we should select the model with the smallest Bayesian Information Criterion (MBIC), Akaike Information Criterion (MAIC), and Hannan-Quinn Information Criterion (MQIC). In general, the first-order panel VAR is the preferred model since it has the smallest MBIC and MQIC. We also check the stability condition of the estimated panel VAR, and we find that all eigenvalues lie inside the unit circle, which means that the estimate is stable. Finally, we check the Hansen's J test, and the

Hansen's J statistics are insignificant, which means that we cannot reject our GMM-style instruments are valid.

The estimation results for our panel VAR model are shown in Table 6. Our main objective is to examine how the team number and kernel number jointly influence the performance of the best solution. The one-period lagged dependent variables allow us to interpret the short-term effect more easily. The results indicate that attracting more teams and encouraging teams to share more solutions at time t-1 positively affects the best performance of all teams in the next period. However, the number of shared solutions negatively moderates the relationship between the team number and the best performance of all teams (i.e., the parallel path effect). These results support our H3a, H3b, and H3c.

Table 6 Panel VAR Estimation for $MaxPrivateScore_{ij}$

	Dependent Variable			
Independent Variable	$\ln(TeamNum_{jt} + 1)$	$\ln(KernelNum_{jt} + 1)$	$\ln(TeamNum_{jt} + 1)$ $\times \ln(KernelNum_{jt} + 1)$	$MaxPrivateScore_{ij}$
$\ln(TeamNum_{jt-1} + 1)$	.9363887*** (.0804128)	.0132983 (.0517201)	.5886496 (.7945301)	.4258699*** (.1504277)
$\ln(KernelNum_{jt-1} + 1)$	.394379** (.1714858)	1.008163*** (.1167329)	2.797244 (1.702409)	.8511826** (.3406379)
$\ln(TeamNum_{jt-1} + 1)$ $\times \ln(KernelNum_{jt-1} + 1)$	-.0295421** (-.0295421)	-.0103772 (.0089431)	.6671285*** (.1151961)	-.0673906** (.0279781)
$MaxPrivateScore_{ij-1}$	-.0615889* (.0324364)	-.0164572 (.0321962)	-.5550865* (.335936)	.6563923*** (.1023322)

**** p<0.01, ** p<0.05, * p<0.1*

5. Discussion and Conclusion

Our studies theoretically proposed and empirically examined how the contestants with different skill levels are influenced differently by the exposure to shared solutions through learning and fixing effects. Our results suggest that the overall effects of exposure to shared solutions on the contestants are favorable in general, but high-skilled contestants benefit less on average. However, the effects on the crowd's best performance are not necessarily positive. Even though exposure to shared solutions has a positive main effect on the performance, it may also discourage parallel path effects.

This work offers both theoretical and managerial contributions. Theoretically, our study adds to the crowdsourcing literature by providing a detailed analysis of the effect of allowing solution sharing on different stakeholders in the crowdsourcing contest platform. Research regarding the effects of exposure to solutions generated and shared by other solvers in crowdsourcing contests remains nascent and underexplored ([29] - [32] are a few exceptions). Current empirical studies mainly studying this effect focus more on how different dimensions of a shared solution (e.g., originality, quality) influence the contestants' performance (Ba et al. 2017, Jin 2018) [31]-[32]. Our study focuses on how contestant's quality moderates the effect of exposure to shared solutions on the performance. We point out that the sharing may lead to unintended outcomes (e.g., reduced parallel path effect).

From the managerial perspective, this work offers insights for managers who are currently debating the legality of allowing solution sharing in the crowdsourcing contest platform for data science tasks. Even though allowing solution sharing might be beneficial to the platform, the requesters, and the contestants on average, managers in the crowdsourcing contest platforms still need to be cautious when they make a decision related to allowing solution sharing because of the existence of unintended adverse effects. The decision should be made based on the tradeoff between the benefits (e.g., learning effect) and the costs (e.g., fixation effect) of the solution sharing. For example, if some contests want to make new contestants familiar with the task quickly, motivating more shared solutions is the dominant strategy. However, if some contests already have attracted a large number of teams from diverse backgrounds, discouraging solution sharing (at least the low-quality solutions) might be the dominant strategy, given shared solutions may inhibit the parallel path effect.

Several future extensions are possible. In this study, we could not observe how and to what extent contestants used the shared solutions based on our data, so we use voting behavior as a proxy for the exposure to the shared solutions. Future studies focusing on the effects of exposure to shared solutions in the crowdsourcing contest may want to measure solution usage more directly with the proper private dataset (e.g., the log data). For instance, with the log data, scholars can measure how much time contestants spent on exploiting each existing solution and how much effort contestants spend. Second, we only examined the effect of solution sharing for the data-science related tasks. Future studies can also examine the impact on other types of crowdsourcing tasks, including the tasks evaluated by the subjective criteria (e.g., the logo design tasks and web development tasks). Third, even though we examined how contestants' skill level moderates the effect of shared solutions, future studies can focus more on how the types (e.g., educational purpose versus solution leaking purpose) and the quality (e.g., high- versus low-quality) of shared solutions moderate the effect. Forth, although we use the fixed effects as our main identification strategies to examine how exposure to shared solutions influences contestants' performance (a common identification strategy used for the online crow platform studies, e.g., [33]-[34]), it is still interesting to test the predictive power of these shared solutions using machine learning techniques.

Appendix. Robustness Check

A1. An Alternative Measurement of Team Performance

For our H1, we use the best daily performance instead of average daily performance to measure the new entrant teams' performance. The results are highly consistent.

For our H1, H2 and H3, we use the performance in the private leaderboard as the measurement of teams' performance. As a robustness check, we use the performance in the public leaderboard as the measurement of teams' performance. All results using alternative measurements are consistent.

A2. Heteroscedasticity-based Instrument

To examine whether the shared solutions help new entrant teams to achieve better performance at their first submissions (i.e., our H1a), we use the mathematically generated instrumental variables to test the robustness of our OLS estimates. We construct the orthogonal instruments mathematically using the method proposed by Lewbel [35]. Lewbel's method treats all covariates as exogenous and constructs orthogonal instruments mathematically from these covariates (e.g., [36]). In our model, we treat the quality of new entrants and duration elapsed as exogenous variables to construct instruments for our endogenous term (i.e., the shared solutions number at time t-1). We implement this method using xtivreg2h command in STATA, and we account for the contest-specific fixed effects using the fe option. The highly consistent results indicate that the specific instruments we chose do not drive our 2SLS results.

Table 7 the Impact of Solution Sharing on New Entrant Teams' Performance (Lewbel type IV)

Model	(1)	(2)	(3)	(4)
DV	$AvgPublic$	$AvgPrivate$	$MaxPublic$	$MaxPrivate$
Estimator	Generated IVs			
ln($KernelNum$ 1)	0.137***	0.144***	0.252***	0.247***
	(0.0402)	(0.0403)	(0.0386)	(0.0388)
ln($NewUser$ + 1)	-0.414***	-0.490***		
	(0.0747)	(0.0751)		
ln($AvgRanking$ 1)	-1.223***	-1.168***	-0.940***	-0.923***
	(0.101)	(0.104)	(0.0934)	(0.0954)
$TimeElapse$	0.266***	0.238***	0.702***	0.622***
	(0.0619)	(0.0620)	(0.0542)	(0.0544)
Observations	10,227	10,104	10,399	10,310
R-squared	0.041	0.039	0.150	0.127
Contest FE	YES	YES	YES	YES
Kleibergen-Paap rk LM statistic	99.312	102.018	103.390	103.059
Cragg-Donald Wald F statistic	1252.981	1303.022	1232.760	1220.444
Hansen J statistic	0.912	3.080	4.626	3.035
P-value for Hansen J statistic	0.6339	0.2144	0.0990	0.2192

*Notes. Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1*

A3. Alternative Measurement of Exposure to Shared Solutions at the Team Level

For our team-level analysis, we use the solution voting behavior to measure contestants' exposure to shared solutions. Even though it is the platform norm for contestants to give votes to the shared solutions they used, it is still possible that some contestants used the shared solutions without voting. To mitigate this concern, we use the number of shared solutions available in the last period to measure the team's exposure to shared solutions. We use this contest-level measurement as a robustness check for our H2, and our results are highly consistent.

7. References

[1] Boudreau, K. J., & Lakhani, K. R. (2013). Using the crowd as an innovation partner. Harvard business review, 91(4), 60-9.

[2] Erat, S., & Krishnan, V. (2012). Managing delegated search over design spaces. Management Science, 58(3), 606-623.

[3] Stanko, M. A. (2016). Toward a theory of remixing in online innovation communities. Information Systems Research, 27(4), 773-791.

[4] Bandura, A., & Walters, R. H. (1977). Social learning theory (Vol. 1). Englewood Cliffs, NJ: Prentice-hall.

[5] Kolb DA (1984) Experiential Learning: Experience as the Source of Learning and Development (Prentice-Hall, Englewood Cliffs, NJ).

[6] March JG (1991) Exploration and exploitation in organizational learning. Organ. Sci. 2(1):71–87.

[7] Levinthal, D. A., & March, J. G. (1993). The myopia of learning. Strategic management journal, 14(S2), 95-112.

[8] Youmans RJ, Arciszewski T (2014) Design fixation: A cloak of many colors. Gero JS, ed. Design Computing and Cognition 12, 115–129.

[9] Wang, K., Nickerson, J., & Sakamoto, Y. (2018). Crowdsourced idea generation: The effect of exposure to an original idea. Creativity and Innovation Management, 27(2), 196-208.

[10] Koh, T. K. (2019). Adopting seekers' solution exemplars in crowdsourcing ideation contests: antecedents and consequences. Information Systems Research, 30(2), 486-506.

[11] 'O'Cass, A., Heirati, N., & Ngo, L. V. (2014). Achieving new product success via the synchronization of exploration and exploitation across multiple levels and functional areas. Industrial Marketing Management, 43(5), 862-872.

[12] Jansson, D. G., & Smith, S. M. (1991). Design fixation. Design studies, 12(1), 3-11.

[13] Smith, S. M. (2003). The constraining effects of initial ideas. Group creativity: Innovation through collaboration, 15-31.

[14] Kohn, N. W., & Smith, S. M. (2011). Collaborative fixation: Effects of 'others' ideas on brainstorming. Applied Cognitive Psychology, 25(3), 359-371.

[15] Smith, S. M., Ward, T. B., & Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. Memory & cognition, 21(6), 837-845.

[16] Nelson, R. R. (1961). Uncertainty, learning, and the economics of parallel research and development efforts. the Review of Economics and Statistics, 351-364.

[17] Abernathy, W. J., & Rosenbloom, R. S. (1969). Parallel strategies in development projects. Management Science, 15(10), B-486.

[18] Dahan, E., & Mendelson, H. (2001). An extreme-value model of concept testing. Management science, 47(1), 102-116.

[19] Sommer, S. C., & Loch, C. H. (2004). Selectionism and learning in projects with complexity and unforeseeable uncertainty. Management science, 50(10), 1334-1347.

[20] Terwiesch, C., & Xu, Y. (2008). Innovation contests, open innovation, and multiagent problem solving. Management science, 54(9), 1529-1543.

[21] Boudreau, K. J., Lacetera, N., & Lakhani, K. R. (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. Management science, 57(5), 843-863.

[22] Jeppesen, L. B., & Lakhani, K. R. (2010). Marginality and problem-solving effectiveness in broadcast search. Organization science, 21(5), 1016-1033.

[23] ' O'Leary, D. E. (2019). An empirical analysis of information search and information sharing in crowdsourcing data analytic contests. Decision Support Systems, 120, 1-13.

[24] Dissanayake, I., Zhang, J., Yasar, M., & Nerur, S. P. (2018). Strategic effort allocation in online innovation tournaments. Information & Management, 55(3), 396-406.

[25] Liu, T. X., Yang, J., Adamic, L. A., & Chen, Y. (2014). Crowdsourcing with all-pay auctions: A field experiment on taskcn. Management Science, 60(8), 2020-2037.

[26] Machado, J. A., & Silva, J. S. (2019). Quantiles via moments. Journal of Econometrics, 213(1), 145-173.

[27] Abrigo, M. R., & Love, I. (2016). Estimation of panel vector autoregression in Stata. The Stata Journal, 16(3), 778-804.

[28] Andrews, D. W., & Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. Journal of econometrics, 101(1), 123-164.

[29] Wooten, J. O., & Ulrich, K. T. (2015). The impact of visibility in innovation tournaments: Evidence from field experiments. Available at SSRN 2214952.

[30] Majchrzak, A., & Malhotra, A. (2016). Effect of knowledge-sharing trajectories on innovative outcomes in temporary online crowds. Information Systems Research, 27(4), 685-703.

[31] Ba, S., Jin, Y., Lee, B., & Stallaert, J. Winning by Learning?.'

[32] Jin, Y. (2018). Essays on Online Two-sided Platforms.

[33] Sabzehzar, A., Hong, Y., Burtch, G., & Raghu, T. S. (2019). The role of religion in online prosocial lending. In 40th International Conference on Information Systems, ICIS 2019 (40th International Conference on Information Systems, ICIS 2019). Association for Information Systems.

[34] Sabzehzar, Amin and Burtch, Gordon and Hong, Yili and Raghu, T. S., The Role of Religion in Online Pro-social Lending: An Interactional View (June 20, 2020). Available at SSRN: https://ssrn.com/abstract=3631704 or http://dx.doi.org/10.2139/ssrn.3631704