Collaborate to Compete: An Empirical Matching Game under Incomplete Information in Rank-Order Tournaments

Tat Chan, Yijun Chen and Chunhua Wu*

October 2022

Abstract

This paper studies the collaboration of talents in rank-order tournaments. We use a structural matching model with unobserved transfers among participants to capture the differentiated incentives of participants that spur collaboration, with a specific focus on incorporating incomplete information and competition in the matching game. We estimate our model using data from a leading data science competition platform and recover the heterogeneous preferences and abilities of participants that determine whether and with whom they form teams. Overall, teamwork enhances performance and competition fosters collaboration, whereas incomplete information about potential coworkers' ability hinders collaboration. Using the estimation results, we conduct counterfactuals to investigate how the information on potential collaborators' ability and competitive pressure affect collaboration and performance outcome. Our results suggest that the platform could further improve collaboration and yield better outcomes by providing more informative signals of ability and further concentrating the allocation of rewards to top performers.

Keywords: Matching game, Collaboration, Incomplete information

*Tat Chan (chan@wustl.edu) is Professor of Marketing at the Olin Business School, Washington University in St. Louis. Yijun Chen (yijun.chen@imperial.ac.uk) is Assistant Professor at the Imperial College Business School, Imperial College London. Chunhua Wu (chunhua.wu@sauder.ubc.ca) is Associate Professor of Marketing and Behavioral Science at the Sauder School of Business, University of British Columbia. The authors thank seminar participants at Washington University and the 2017 Marketing Science Conference for their valuable feedbacks.

1 Introduction

Collaboration enhances the performance of individuals and organizations in various contexts (e.g., Hollis, 2001; Ductor, 2015; Un et al., 2010). Competition often fosters collaboration. For example, more than 200 COVID-19 vaccine candidates were in development in 2020. In the race to become the first to succeed, at least one-third of the projects were partnerships (Druedahl et al., 2021). Since collaborations require cooperation and/or specialization of complicated tasks, their success relies on the ability of all parties. A lack of information about potential partners would lead to greater uncertainty in outcomes and thus deter collaboration.

This paper empirically studies how information availability and competitive pressure affect collaborations and their outcomes, focusing on a unique setting: crowdsourcing data science competitions on Kaggle.com. Kaggle is a leading platform connecting firms sponsoring competitions with participants (i.e., data scientists) who compete to develop the best solutions and win rewards. The rewards are both monetary (cash prizes to top teams offered by sponsoring firms) and non-monetary (Kaggle points rewarded by the platform) and are based on performance rankings–a format equivalent to rank-order tournaments. A competition typically attracts hundreds or even thousands of participants, who are encouraged to collaborate as teams. There are multiple reasons for individuals to form teams. First, collaboration can bring many benefits, including the economy of scale, complementarity of knowledge and skills, and division of labor, all of which help tackle complicated tasks and improve overall performance. Competitive pressure is another reason. Because payoffs from the top ranks are much larger than those of other ranks in the competition, the collaboration of other participants may force an individual also to collaborate. However, as repeated interactions across competitions are not common, individuals have to rely on incomplete public information, such as Kaggle tiers[1], to infer the real ability of potential teammates when forming teams. This can lead to inefficiencies due to adverse selection and moral hazard (e.g., Holmstrom, 1982; von Siemens and Kosfeld, 2014), which lower participants' incentives to form teams. As we show in this paper, collaborations greatly improve the overall outcomes of Kaggle competitions; thus understanding how information and competition affect collaborations will build knowledge of how a better platform design can motivate collaborations and improve team performances.

To achieve that goal, we develop a structural one-sided matching model to quantify participants' underlying motives for collaboration. The model has several unique features

[1]Source: https://www.kaggle.com/progression. The Kaggle tiers are part of the progression system that is based on participants' past performances. See Section 3 for more details.

that capture the empirical setting. First, participants compete against each other in the market. A successful collaboration will improve their ranking and thus the monetary and non-monetary rewards while at the same time reducing the payoffs of other teams. Second, the matching game accommodates a large number of participants with heterogeneous abilities, which are partially reflected by incomplete public information. Third, our model allows potential collaborators to negotiate how rewards and costs are shared. Teams will be formed only if all the parties agree on the sharing rule. These features allow us to capture participants' decisions to collaborate in the competition under incomplete information. We focus on the market equilibrium, which is characterized by each participant's optimal choice regarding whether to collaborate and with whom, under two constraints. The first is that each individual makes rational inferences on the true ability of potential teammates based on their public information and collaboration decisions. The second constraint requires that the sharing of rewards and costs agreed upon by collaborators clears the market.

Estimation results show that a participant's tier status reasonably reflects her actual ability. However, there is considerable variation in abilities across participants who belong to the same tier, indicating that the public information is a noisy signal. We find that both monetary and non-monetary rewards, including Kaggle points and other benefits from forming teams, are highly valued by participants. We also find that participants, in general, perform much better by forming teams. However, for high-ability participants, the gains from collaborating with lower-ability teammates are negative, implying the risk of collaboration due to incomplete information. Finally, we recover the market-clearing reward-sharing rules between participants under market equilibrium. Results show that participants at a lower tier are willing to pay a positive (monetary and non-monetary) transfer to teammates at a higher tier. This explains why a significant proportion of participants from high tiers are still willing to form teams with participants from lower tiers in the data.

Based on the recovered model primitives, we conduct two counterfactuals to investigate how to improve platform design to facilitate collaborations and enhance team performances. In the first counterfactual, we show that uncertainty about the ability of potential teammates is an important factor that hinders collaboration among high-ability participants. Increasing the informativeness of the tier status could reduce the uncertainty and lead to more collaborations among high-ability participants. It will boost the performance of the best teams, but at the cost of the performance of an average participant who may find it more difficult to team up with high-ability teammates. In the second counterfactual, we manipulate the competitive pressure by changing how the platform allocates its Kaggle points to teams at different performance rankings. We find that increasing the competitive pressure from a flat scheme of points allocation to a steep scheme that concentrates the points' allocation to

top teams will incentivize all participants to collaborate. Consequently, more teams will be formed and both the best and average team performance will be improved, and this creates benefits for the sponsoring business and the participants.

The contribution of this paper is two-fold. From the methodology perspective, we develop an empirical matching model that explicitly accounts for incomplete information and competition—two factors not addressed in the extant empirical matching literature (e.g., Choo and Siow, 2006; Fox, 2008; Sørensen, 2007). Our model also allows for unobserved sharing rules between agents, a factor not accounted for in the previous empirical literature of coalition games. Our utility-based model framework is suitable for estimating large matching games with incomplete information and/or externality, such as matching in online dating platforms and team formation in multiplayer games. For substantive contributions, this paper provides insights on how collaborations affect individual performance and competition outcomes, and how information and competition affect team formation and performance. The results contribute valuable insights on efficient platform designs in the environments under which collaborations are prevalent.

The rest of the paper is organized as follows. We discuss related literature in Section 2, then describe the empirical context and provide summary statistics in Section 3. Section 4 develops the matching model. Detailed model specification, identification and the estimation are presented in Section 5, followed by estimation results in Section 6. Counterfactual analyses are shown in Section 7. Finally, we discuss the limitations and outline future research directions in the conclusion section.

2 Related Literature

Our study is closely related to the large stream of literature on matching. Theoretical works on matching games have been developed for decades. The "Gale-Shapley" algorithm has been applied to solve problems for college admissions (Gale and Shapley, 1962), dating markets (Becker, 1973), and business and plant locations (Bayus, 2013). While most of the works assume complete information in the matching game, a few recent papers explore the properties of the matching game when agents have incomplete information. Liu et al. (2014), for example, study a matching game with one-sided incomplete information and show that the set of stable outcomes is nonempty and is a superset for the set of complete-information stable outcomes.

Empirical works on matching are rather recent. Fox (2008) proposes using the maximum score estimator to estimate the matching game. In a later paper (Fox, 2010), he discusses the identification conditions for using observed matching outcomes for model estimation. The

maximum score estimator has been applied in several studies in different industries (e.g., Fox and Bajari, 2013; Yang et al., 2009; Wu, 2015; Ni and Srinivasan, 2015). A few recent papers study the vertical relationship between insurance networks and hospitals using the matching model. Ho and Lee (2017), for example, use a Nash-in-Nash framework as the equilibrium concept in the matching game. A similar modeling approach is adopted in Ghili (2022). Our matching model assumes that there is a sharing rule (under which each party is making the optimal choice in matching and the market is cleared) between collaborators depending on their attributes. This approach was first developed in Becker (1973) and later adopted in the empirical work of Choo and Siow (2006) that studies the marriage market. These two papers, as well as other empirical studies mentioned above, do not consider the issue of incomplete information. In this sense, our paper is close to Chan et al. (2015), who use a matching model to study how individuals, fully aware of the costs associated with being infected, engage in risky sex behaviors. Agents in their model have uncertainty regarding the health status of their partners. Chan et al. (2015) also make the market-clearing assumption so that they can estimate the model using the maximum likelihood estimator with equilibrium constraints. Our model diverges from theirs by incorporating competition among collaborations, under which the payoff of one collaboration is affected by the performance of the others.[2]

Collaborations are typically modeled as a coalition game (for example, see Farrell and Scotchmer, 1988; Pycia, 2012). An agent's payoff is usually assumed to be determined by the coalition she belongs to. In a more complex setting, the payoff can be determined by other coalitions, and the agent thus will react to other agents' coalition decisions under the competitive pressure (Yi, 1997; Wilson et al., 2010). Our study fits into the framework of a coalition game when externality exists. We contribute to this stream of literature by relaxing the perfect information assumption and allowing for an unobserved sharing rule for the coalition formation. Our study incorporates externality by directly modeling the payoff of a coalition as a function of other coalitions.

The empirical context of our paper is aligned with the growing literature on crowdsourcing. Given the emergence of crowdsourcing platforms in the past decade, researchers have explored various phenomena in crowdsourcing. Burtch et al. (2013), for example, study the content contribution of users of a digital journal and test several economic theories using substitution models and reinforcement models. Bayus (2013) studies individual ideators'

[2]A few other empirical studies consider either incomplete information or competition. Ackerberg and Botticini (2002), for example, relax the perfect information assumption and estimate contract determinants by explicitly embedding endogenous selection in the matching process. Wilson et al. (2010) extend the matching literature by incorporating externalities from network effects in faculty members' office choice. Uetake and Watanabe (2017) study firm entry decisions in the banking industry, allowing for potential spillovers. The modeling approaches in these papers differ from ours.

contribution in Dell's IdeaStorm community over time, and finds that past success has a negative effect on the current contribution. Huang et al. (2014) study the learning process of participants on the same Dell platform and show that individuals learn quickly about their ability to generate high-potential ideas but are relatively slow to learn the cost of implementation. The above research on crowdsourcing has treated collaborations on platforms as exogeneously given. We contribute to the literature by studying how participants in Kaggle competitions collaborate and how collaboration affects the outcomes.

Finally, the way that Kaggle awards participants Kaggle points and monetary prizes makes the competitions equivalent to rank-order tournaments (Lazear and Rosen, 1981). This stream of literature studies how ranking-based rewards can motivate hard work and improve performance. For example, Eriksson (1999) uses compensations for executives to test the tournament theory. Kini and Williams (2012) find that higher tournament incentives will motivate risk-taking behaviors for senior managers in order to increase the chance of being promoted. Lazear (1989) shows that while tournaments motivate worker effort, excessive competition for rewards may reduce collaborations. Our study differs from these works by investigating how different competitive pressure driven by the structure of rewards (Kaggle points) affects the collaboration and performance outcomes.

3 Background and Data

In this section, we discuss the empirical context, describe the data, and explore some data patterns that are related to our empirical matching model.

3.1 Empirical Setting

Our empirical setting is Kaggle.com, a leading global crowdsourcing platform for predictive modeling and analytics competitions. Founded in 2010, Kaggle bridges the connection between the demand for and supply of data science talent. On the demand side, firms provide data for the business problems they seek to solve or opportunities they want to explore. On the supply side, data scientists, researchers, and students who have the talent and tools to solve the problems crave for the opportunity to prove their ability and earn rewards. Kaggle connects the two sides by holding sponsored crowdsourcing competitions in which participants compete to provide the best solutions and win awards offered by sponsoring businesses. By the end of 2017, Kaggle had hosted 248 competitions, attracted more than 60,000 participants, and awarded over 9 million US dollars. These competitions have resulted in significant scientific advancements including furthering the state of the art in HIV research, improv-

ing predictive technologies and algorithms, and uplifting operational efficiency in business applications.[3]

The sponsoring business specifies the winning rules and monetary prizes for most competitions. Depending on the business background, type of analytics required, and the winning prize, each competition attracts a distinct set of participants. They can compete solo or form teams. It is common for thousands of participants to register in the same competition, making it very difficult to win the prize. To incentivize participation, Kaggle awards "Kaggle points" to each participant based on her final ranking in the competition. The typical policy is to allocate most points to a few best performers; then, the awarded points decline quickly as a convex function with lower ranks. This creates additional competitive pressure for participants to be placed high in competitions when the Kaggle points are valuable. We capture the competitive nature in our structural matching model.

Kaggle uses a tier progression system to classify individuals. The tier system is based on participants' past performances on the platform. Participants who are new to the platform, thus with no previous performances, are recognized as *Novices*. After gaining points in one competition, participants progress to the *Contributor* tier. To further progress to the *Expert* tier, participants need to consistently achieve good performances (e.g., top 40% of all the teams) across competitions. Finally, the highest tier on Kaggle, *Master*, requires one to consistently place at the very top (e.g., top 10% of all the teams) across competitions.[4] In addition to Kaggle tiers, Kaggle has a leaderboard that ranks participants based on accumulated points across competitions. Note that the amount of earned points decays over time; thus to rank high on the Kaggle leaderboard, participants need to continuously participate and place high in competitions. As Kaggle has gradually established its reputation in the data science community, showing the top-tier status and a high ranking on Kaggle's leaderboard is a useful way to strengthen the resume of data scientists. In an interview with Wired Magazine, Gilberto Titericz, a top Kaggle player, shared that job opportunities that flow from a good Kaggle ranking are generally more bankable than money prizes.[5]

Collaborations can be critical for participants to achieve good performance and win rewards. To ensure high-quality solutions for sponsoring businesses, Kaggle designs rules that do not simply allow, but encourage participants to collaborate. Indeed, the formal unit of participation is a "team," such that a participant competing solo is just a "single-

[3]Source: https://techcrunch.com/2017/06/22/the-kaggle-data-science-community-is-competing-to-improve-airport-security-with-ai and https://www.kaggle.com/c/passenger-screening-algorithm-challenge.

[4]See https://www.kaggle.com/progression for a more detailed description on the point allocation and tier progression system. Some terminologies has changed in 2016. We use the ones before the change in this paper.

[5]Source: https://www.wired.com/story/solve-these-tough-data-problems-and-watch-job-offers-roll-in.

member team." Collaborations on Kaggle are formed in a decentralized way, in which participants decide whether and with whom to form teams, usually through an invitation-and-acceptance/rejection procedure. Mutual agreements among team members are needed, but once the team is formed, a participant is not allowed to separate from the team throughout the competition.

Despite the potential benefits of collaboration, many factors may deter team formation. The first is the lack of information about other participants' actual ability. Kaggle tries to solve this problem by making each participant's tier-status information public on the website. This information, nevertheless, is an imperfect measure. For example, the abilities of Novices are not well distinguished, and individuals who participated in more competitions are more likely to belong to a higher tier. [6] Furthermore, how the monetary and non-monetary rewards are shared among team members may also affect collaborations. The monetary prize—if a team wins—has to be split between members with rules negotiated in advance. Kaggle points will also be allocated based on the number of participants in a team (more details are below). Finally, free-riding and moral hazard can create inefficiencies and conflicts among team members. Therefore, the expected payoff for a participant in a team may not be higher than that from competing solo.

To encourage collaboration, Kaggle changed its point-allocation policy in 2016. Before the change, team members received points according to what their team won divided by the team size. The new policy divides the points of the team by the square root of the team size.[7] Kaggle also reduced the points a team can win at each rank. Single-member teams, therefore, can earn fewer points under the new policy. This could increase the incentive for participants to form teams.

3.2 Data and Summary Statistics

We use the Meta Kaggle data provided by Kaggle.com[8] for the empirical application. The dataset includes information on competition rules, participants, teams, and the final scores, ranking, and monetary rewards for each team. We observe 315 competitions that cover a period of 7 years, from 2010 to 2016. We exclude competitions that did not award Kaggle points, which are educational or designed for participants to familiarize themselves with the

[6] It's possible that some teams are formed based on previous interactions that reveal better information than Kaggle tiers, such as students from the same university or players who have collaborated on Kaggle before. Overall, the percentage of such teams is small: our manual check shows that less than 3% of teams use names revealing the university name, and less than 3% of teams have the same team members in multiple competitions in our sample.

[7] Source: https://blog.kaggle.com/2015/05/13/improved-kaggle-rankings.

[8] Source: https://www.kaggle.com/kaggle/meta-kaggle.

platform. We also exclude competitions that have less than 100 participants. These are mostly competitions in the very early years of Kaggle. Since the value of Kaggle points was not yet well recognized by the data scientist community, the incentives for participants could be different from the competitions in a later stage when Kaggle points are highly valued. A competition lasts several months, during which a team makes an average 23 submissions. Only the best submissions are evaluated for ranking and reward. We exclude teams with less than 5 submissions since their incentives to participate could be different from teams that actively make submissions. For example, the participants may only want to taste what Kaggle's competitions are about. After excluding the two types of competitions and participants with too few submissions, we retain 102 competitions and 32,362 unique participants in the model estimation. In the sample, 87% are single-member teams. For teams with multiple members, 63% have two members. The dimensionality of team options will become much higher and the matching problem too complex if we model team formation with more than two members. For the simplicity of analysis, we assume that, for teams with more than two members, the formation is driven by multiple, separate one-on-one matching between the member with the highest cumulative Kaggle points and each of the other members. The rationale for this assumption is that the presence of the member with the highest tier status is the most important determinant for the team performance. For instance, the performance of teams whose highest tier is Master has first-order statistical dominance over teams whose highest tier is below the Master level.

Table 1 presents summary statistics on quartiles of monetary rewards and number of participants across competitions. The monetary rewards vary dramatically across competitions: the average prize of the first quartile is only $380, but it increases to $67,621 in the fourth quartile. In general, higher monetary rewards attract more participants. Competitions in the first quartile of monetary rewards on average attract 226 participants, while those in the fourth quartile attract 1,374 participants.

Table 1: Monetary Rewards and Participants Across Competitions

Rewards Quartile	Rewards (USD)			Participants		
	min	mean	max	min	mean	max
Q1	250	2319	5000	107	380	1840
Q2	6000	9054	10000	108	545	3848
Q3	13000	20345	25000	194	896	2362
Q4	30000	88500	500000	196	1721	5696

Table 2 shows Kaggle points of the four tiers. We report the average points a participant wins from each competition she joined. Novices have not participated in any competition and

thus have 0 points. As a participant's tier moves up, the average Kaggle points received in each competition also increase. The last column of the table shows that 60% of participants are Novices, whereas Masters are an elite group, as only 10% of participants belong to this tier.

Table 2: Summary Statistics of Participant Tiers and Kaggle Points

Tier	Mean Points per Competition	No. of Participants
Novice	0	34929 (60.0%)
Contributor	845	10924 (18.7%)
Expert	1702	6679 (11.5%)
Master	4371	5721 (9.8%)

Note: Participants may join multiple competitions, so the total number of participants in this table is larger than 31,246(unique participants across competitions); Percentages of each tier are shown in parentheses.

Table 3 reports how participants with different tiers choose to form teams. Two clear patterns arise. First, Novices and Masters are more likely to form teams, likely due to different reasons. Novices form teams in order to learn and improve their ability by collaborating with others. Masters, on the other hand, are well-recognized in the community for their high abilities. Since they can help to increase the team performance, Masters are highly sought after for collaborations. Second, there is a pattern of sorting: participants tend to match with other participants from the same tier. This is especially true for both Novice and Master tiers. The proportion of teams formed with Novices is large across tiers because Novices are the majority in most competitions.

Table 3: Summary Statistics of Participant Team Choices

Tier/Choice	Single	Team			
		Novice	Contributor	Expert	Master
Novice	54%	39.7%	3.2%	1.4%	1.6%
Contributor	78.0%	10.8%	8.3%	1.9%	0.9%
Expert	75.6%	7.5%	3.0%	9.8%	4.1%
Master	53.2%	8.7%	1.4%	4.1%	32.5%

Note: Rows represent participant tier and columns represent participants' choices. Numbers represent percentage of choices for each option.

We now look at how collaboration impacts performance. In almost all the competitions, performance is measured by the predictive accuracy on hold-out samples, but the criteria used for calculating the accuracy differ from competition to competition.[9] Since the measure

[9]Some of the most commonly used evaluation algorithms are Root Mean Squared Errors (RMSE), Root

is unique for each competition, we create a standardized score from the original performance measure. We first calculate the mean and standard deviation of the original performance measure (i.e., accuracy metric) for all single-member teams in the data, including those teams with less than 5 submissions. We then deduct the original performance of each team by the calculated mean and further divide it by the calculated standard deviation. The idea of the standardization is that the mean and standard deviation of single-member teams capture the benchmark difficulty and variation in performance using the original measure. After the standardization, the scores of all teams can be compared across competitions.

We report the average and the standard deviation of the standardized scores across different types of teams in Table 4. Several interesting patterns arise. First, conditional on a participant's tier, the performance when forming a team is in general better than when one competes solo. Since the performance provides value for sponsoring businesses, this result suggests that Kaggle should encourage more collaborations, a direction it has long pursued. Still, a large percentage of participants compete solo, as shown in Table 3. One major reason that we capture in the model is that, since participants have to split the monetary prize and Kaggle points and face potential team conflicts, the expected payoff of a team player could be lower than if she competes solo. Second, teaming with a higher-tier participant improves performance—the average performance is the highest when working with a Master. Finally, standard deviations in parentheses reveal a large variation in the performance of each team type. Use single-member Novice teams as an example: the top quartile score higher than 1.03, while the bottom quartile score lower than 0.26. This implies that the ability of participants who belong to the same tier is highly heterogeneous. Therefore, it is important to capture the uncertainty about a potential teammate's true ability in the matching model, which we will present in the next section.

4 Model

In this section we develop a structural matching model, explicitly incorporating incomplete information and competition, to study the matching outcomes when the market is at equilibrium.

We model a participant's team formation decision as a one-sided matching game in a market with a large number of individuals. In Figure 1, we illustrate the empirical process with a flow chart. The first box on the left represents a participant entering a competition with her tier (Novice, Contributor, Expert, or Master) that is public information, and her

Mean Squared Errors (RMSE), Root Mean Squared Logarithmic Error (RMSLE), Area Under Receiver Operating Characteristic Curve (AUC), and Log Loss.

Table 4: Summary Statistics on Team Type and Performance Outcomes

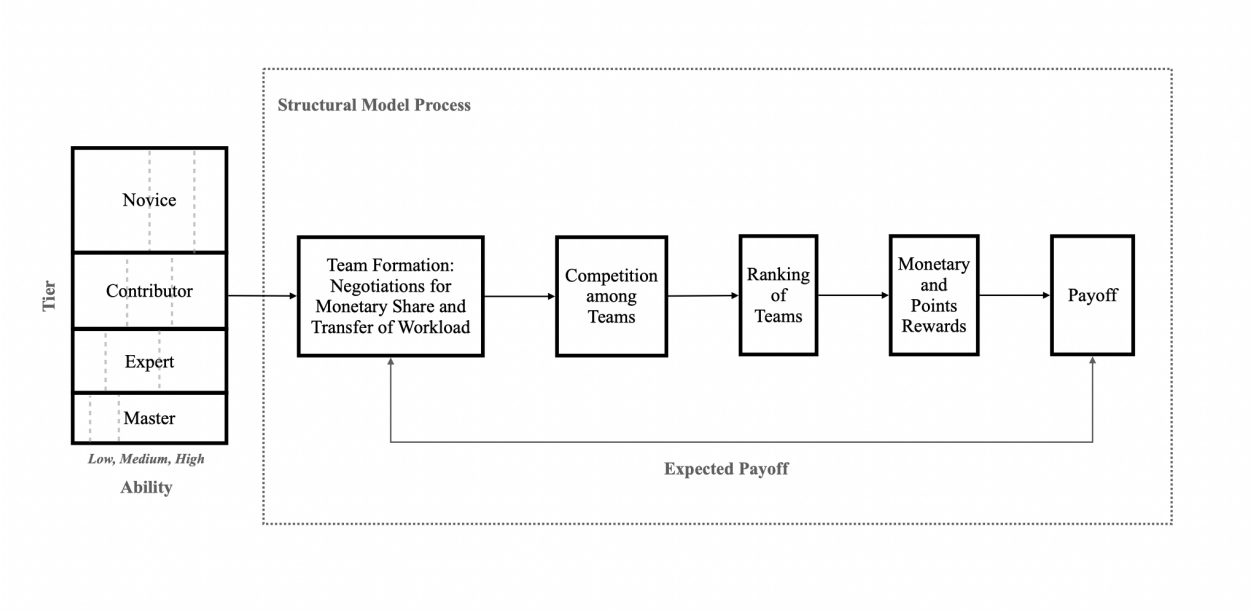
Tier/Choice	Single	Team			
		Novice	Contributor	Expert	Master
Novice	0.64	0.77	0.76	0.95	1.32
	(0.79)	(2.7)	(0.74)	(0.87)	(0.89)
Contributor	0.61		0.83	0.95	1.07
	(0.61)		(0.75)	(1.07)	(0.76)
Expert	0.82			1.13	1.19
	(0.62)			(0.89)	(0.81)
Master	1.03				1.43
	(0.83)				(0.90)

Note: Each row represents a participant's tier and each column represents the participant's team choice. Each number represents the mean score for a team type, and the standard deviation is in parentheses. The statistics are computed for teams with at least 5 submissions in the competition.

true ability (low, medium, or high) that is private information. The following boxes show the processes modeled in this study. The second box from the left indicates that participants will first negotiate with potential teammates on the share of the monetary reward (if any) and the "transfer" of workload from one to another. Next, teams will compete on developing the best solutions (the third box) and will be ranked based on their submissions (the fourth box). Consequently, the monetary and points rewards will be allocated to each team according to the performance ranks (the fifth box) and each participant will obtain her payoff determined by the team rewards, the negotiated shares, and the amount of workload (the sixth box). The arrow from the payoff stage to the team-formation stage indicates that a participant's decision of whether and with whom to form a team is based on her expected payoff from forming a team with each potential teammate in the market. The expected payoff depends on the expected team performance, which is a function of not only the participant's but also the potential teammate's true ability. However, since the participant only observes the potential teammate's tier, she will form a belief on the teammate's true ability based on the tier information, as well as the teammate's decision of forming a team with her under the negotiated terms. We assume that the participant and the potential teammate will form rational beliefs regarding the true ability of each other. We prove in Section 4.4 that the market equilibrium exists in our model.

Below, we formalize the model in four steps. First, we specify the expected payoff function that determines an agent's team-formation decision. We then outline the team-performance function and the monetary and point rewards allocations. We further derive the belief-updating process and the market-clearing conditions. Finally, we prove the existence of the market equilibrium.

Figure 1: Competition Process and Model Structure



4.1 Model Setup and the Payoff Function

We assume that the matching outcomes, including teams that are formed and the performance (i.e., the standardized score) of each team, come from the market equilibrium. To make the model tractable, we make a few additional assumptions. First, the negotiated terms between two parties cannot be broken once the team is formed. Potential issues from forming teams, including moral hazard and personal conflicts that can affect the team performance, are captured in a reduced-form way in the model. Second, the pool of participants is treated as exogenous in the model. This helps us abstract away from the complicated participation problem. To partially address the endogeneity of participation, our model allows the distribution of abilities of participants to be different depending on the monetary reward of competitions. Third, we treat each competition as a static game, so that we can focus on the determinants of team formation within a competition and ignore the strategic dynamic interactions between participants. Fourth, we focus on the team size of two and model only the one-to-one matching decision among participants. In our data, 63% of multi-member teams have a size of two. For teams with more than two members, we model the one-to-one matching decision between the individual with the highest Kaggle tier and each of the other members. Finally, we treat the monetary and non-monetary rewards pre-specified in the competition as exogenous.

For each competition, we denote the set of participants as $\mathcal{N}$, and the number of participants as N. Also denote the set of teams formed as $\mathcal{M}$, and the number of teams as M. A

team $\langle i,j\rangle \in \mathcal{M}$ indicates that the focal participant i forms team with a target participant j. As a special case, $\langle i,\varnothing\rangle$ denotes that participant i competes solo instead of teaming with another individual.

Each participant is represented by two attributes: A_i is the true ability of the participant which is private information, and R_i a noisy signal (i.e., tier status on Kaggle) about her ability that is public information. We assume A_i and R_i are discrete variables, and use A and R to represent the number of possible types for A_i and R_i, respectively. We further use $\mathcal{A}$ and $\mathcal{R}$ to represent the collection of abilities and signals of all participants in the competition. The informativeness of the signal R_i is represented by the conditional probability $Pr\left(A_i = a|R_i\right)$ for all types of abilities. For $a \neq a'$, if $Pr\left(A_i = a|R_i\right)$ is close to $Pr\left(A_i = a'|R_i\right)$, R_i is not informative to identify the true ability. However, if $Pr\left(A_i = a|R_i\right)$ is close to one while the probabilities for other abilities are close to zero, R_i is very informative. The conditional probability is assumed to be common knowledge.

The performance of team $\langle i,j\rangle$, denoted as $Y_{\langle i,j\rangle}(A_i, A_j)$, is determined by the abilities A_i and A_j. Team performance is independent from the ordering of i and j, i.e., $Y_{\langle i,j\rangle}(A_i, A_j) = Y_{\langle j,i\rangle}(A_j, A_i)$, and it does not depend on the abilities of other teams. However, the rank of $Y_{\langle i,j\rangle}$ will depend on the performance of all teams. This captures the competitive environment, as a better performance of one team can drive down the ranking and thus the payoffs for other teams. We use $Y_{\mathcal{M}}$ to denote the collection of performances of all teams under $\mathcal{M}$, and $Z_{\langle i,j\rangle}(Y_{\mathcal{M}})$ to represent the ranking of team $\langle i,j\rangle$. For the sponsoring business, the performance of the best team, i.e., $\max\left(Y_{\mathcal{M}}\right)$, brings the most value as the algorithm can be applied to best solve its business problem. Participants, on the other hand, care about the ranking since it determines the amount of the monetary reward, denoted by $Money(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}))$, and the number of points, denoted by $Point(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}))$, that the team can earn from the competition (details are in Section 4.2).

Kaggle decides how the Kaggle points awarded to a team should be split between its members, and members negotiate by themselves how to share the monetary reward and the team work. Since the abilities A_i and A_j are unobserved by the other team member, the sharing rule will be determined based on the public information R_i and R_j. When the market is at equilibrium, the sharing rule will also be determined by $\mathcal{R}$, the distribution of signals of all participants in the competition (more details are in Section 4.3). Because of this, we use $\gamma^M\left(R_i, R_j, \mathcal{R}\right)$ to represent i's share of the monetary reward. For team $\langle i,j\rangle$, the share of each member is non-negative and the sum of shares is equal to one, i.e., $\gamma^M\left(R_i, R_j, \mathcal{R}\right) + \gamma^M\left(R_j, R_i, \mathcal{R}\right) = 1$. If competing solo, all the monetary reward will belong to the participant, i.e., $\gamma^M\left(R_i, \varnothing, \mathcal{R}\right) = 1$, where "$\varnothing$" indicates that i does not have a teammate. For the share of team work, we use $\tau\left(R_i, R_j, \mathcal{R}\right)$ to denote the "transfer" of

workload from i to j. A larger workload may represent j taking more work tasks or handling work tasks that are less fun or have lower learning value. As a participant's agreement to handle more of the workload implies that the other participant will have less work, we impose the restriction that $\tau(R_i, R_j, \mathcal{R}) + \tau(R_j, R_i, \mathcal{R}) = 0$. This assumption is similar to the model in Becker (1973), in which the transfer of the man and the woman in a marriage sums up to zero. We normalize the transfer between teammates within the same tier to zero, i.e. $\tau(R_i, R_j, \mathcal{R}) = 0$ if $R_i = R_j$. Under the normalization, $\tau(R_i, R_j, \mathcal{R}) > 0$ suggests that j takes more workload, relative to both sharing the work equally. There is no transfer in a single-member team, so $\tau(R_i, \varnothing, \mathcal{R}) = 0$.

Finally, there are additional benefits from team works, including the economy of scale and specialization in job tasks, and as such the workload of each member can be reduced. There are also additional costs, such as moral hazard and potential personal conflicts, when working as a team. Note that these are the benefits and costs on top of how teamwork can impact the performance in the competition, which is captured in $Y_{\langle i,j\rangle}(A_i, A_j)$. We cannot separately identify these additional benefits and costs from the data; therefore, our model only incorporates the net benefit from the above factors. To allow for heterogeneity of the net benefit across teams, we vary it as determined by the tiers of team members, i.e., the net benefit is represented by a function $\alpha(R_i, R_j)$.

Combining the above components, we assume that when the focal participant i is considering the collaboration with the target participant j, she will form an expectation of her payoff relative to all of the other team options. When making the decision, her information set is $(A_i, R_i, R_j, \mathcal{R})$, which also represents the state variables in the expected payoff function. The expected payoff is the following:

$$\begin{aligned} U(A_i, R_i, R_j, \mathcal{R}) =& \theta_i^M \cdot \gamma^M(R_i, R_j, \mathcal{R}) \cdot E\left[Money\left(Z_{\langle i,j\rangle}(Y_{\mathcal{M}})\right) | A_i, R_i, R_j, \mathcal{R}\right] + \\ & \theta_i^P \cdot \gamma^P \cdot E\left[Point\left(Z_{\langle i,j\rangle}(Y_{\mathcal{M}})\right) | A_i, R_i, R_j, \mathcal{R}\right] + \tau(R_i, R_j, \mathcal{R}) + \alpha(R_i, R_j) + \varepsilon_{i,R_j}. \end{aligned} \tag{1}$$

In the above equation, parameters θ_i^M and θ_i^P represent the participant's marginal utility for the monetary and non-monetary reward, respectively. γ^P captures how Kaggle allocates team points to each participant. As discussed in Section 3, under the original policy $\gamma^P = 1/2$, where 2 is the team size; after the policy change, the new $\gamma^P = 1/\sqrt{2}$. Finally, the random component ε_{i,R_j} captures other unobserved factors that will affect the participant decision of whether and with whom she will form a team. We assume that it is the same if two target participants j and j' share the same public signal. That is, $\varepsilon_{i,R_j} = \varepsilon_{i,R_{j'}}$ if

$R_j = R_{j'}$.[10]

Under the normalization assumptions for working solo, the expected payoff function of forming a single-member team is

$$\begin{aligned} U\left(A_i, R_i, \varnothing, \mathcal{R}\right) =& \theta_i^M \cdot E\left[Money\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right)\right)|A_i, R_i, \varnothing, \mathcal{R}\right] + \\ & \theta_i^P \cdot E\left[Point\left(Z_{\langle i,\varnothing\rangle}\left(Y_{\mathcal{M}}\right)\right)|A_i, R_i, \varnothing, \mathcal{R}\right] + \varepsilon_{i,\varnothing}. \end{aligned} \tag{2}$$

The participant will form a team with another participant who belongs to tier R_j if her expected payoff is larger or equal to the expected payoff from forming teams with participants of other tiers (including solo team). Assuming that ε_{i,R_j} follows the Type-I extreme value distribution with scale parameter μ and given a sharing rule γ^M and a transfer rule τ, the probability that participant i's optimal choice is to team up with participant j with signal R_j (including single-member team with $j = \varnothing$ and $R_j = \varnothing$) can be calculated as:

$$Pr(\langle i,j\rangle|A_i, R_i, R_j, \mathcal{R}) = \frac{\exp(V(A_i, R_i, R_j, \mathcal{R})/\mu)}{\sum_{r'\in(R\cup\varnothing)}\exp(V(A_i, R_i, r', \mathcal{R}/\mu)}, \tag{3}$$

where $V(A_i, R_i, R_j, \mathcal{R})$ on the right side is the expected payoff in equation (1)—or equation (2) if $R_j = \varnothing$—without the random component ε_{i,R_j}.

Similarly, the probability that participant j's optimal choice is to team up with participant i with signal R_i is:

$$Pr(\langle j,i\rangle|A_j, R_j, R_i, \mathcal{R}) = \frac{\exp(V(A_j, R_j, R_i, \mathcal{R})/\mu)}{\sum_{r'\in(R\cup\varnothing)}\exp(V(A_j, R_j, r', \mathcal{R}/\mu)}. \tag{4}$$

The two optimal choice probabilities in equations (3) and (4) imply that the matching $\langle i,j\rangle$ (or equivalently $\langle j,i\rangle$) is optimal for both i and j. Note that, γ^M and τ in equation (1) are unobserved by researchers. To evaluate the choice probabilities, we will have to back out them from the model.

In the case of perfect information, i.e. participants' true abilities are common knowledge, the two equations (3) and (4) of optimal choice probabilities could be replaced by:

$$Pr(\langle i,j\rangle|A_i, A_j, \mathcal{A}) = \frac{\exp(V(A_i, A_j, \mathcal{A})/\mu)}{\sum_{a'\in(A\cup\varnothing)}\exp(V(A_i, a', \mathcal{A}/\mu)}, \tag{5}$$

[10]This is based on the assumption that, other than the public signal R_j, the focal participant cannot observe other attributes of the target participant. Therefore, she is indifferent in teaming with j or j' if $R_j = R_{j'}$. Relaxing this assumption makes the matching problem more complicated without direct bearing on our main results. The same assumption is made by Becker (1973), Choo and Siow (2006), and Chan et al. (2015).

and

$$Pr(\langle j,i\rangle|A_j,A_i,\mathcal{A}) = \frac{\exp(V(A_j,A_i,\mathcal{A})/\mu)}{\sum_{a'\in(A\cup\varnothing)}\exp(V(A_j,a',\mathcal{A}/\mu)}, \tag{6}$$

where the true abilities A_i and A_j replace the signals R_i and R_j, and the distribution of abilities $\mathcal{A}$ replaces the distribution of signals $\mathcal{R}$ on both sides. In our application, since participant i only knows R_j in equation (3) and not A_j in equation (5), she will make an inference on A_j based on the public signals, as well as j's decision to form a team with her under the negotiated terms. Details on how she makes the inference are discussed in Section 4.3.

4.2 Performance and Rewards

The ranking of team $\langle i,j\rangle$ depends on the performances of all teams, which are the standardized scores discussed in the previous section, of all teams. With abilities $A_i = a$ and $A_j = a'$, the performance function is specified as

$$Y_{\langle i,j\rangle}(a,a') = \lambda_{aa'} + \xi_{ij}, \tag{7}$$

where $\lambda_{aa'}$ represents the predicted performance of a team with ability a and a' is a model parameter to be estimated. By definition $\lambda_{aa'} = \lambda_{a'a}$.[11] Let $\lambda_{a'}$ represents the predicted performance when j works alone. If $\lambda_{aa'} > \lambda_{a'}$, it means that the performance of working with i is higher than that of no collaboration. It can be due to multiple reasons. One is that i has higher ability than j (i.e. $a > a'$), and j is relying on i to handle challenging tasks. In this case, i's performance does not necessarily improve and could even possibly decrease (i.e., $\lambda_{aa'} < \lambda_a$) because i not only takes the heavy workload but also makes extra effort in communication with j. Another scenario is that complementarity from team work exists, which may come from each team member specializing on the task that she is good at, or simply dealing with less parts of the work can lead to better performance. Complementarity, if it exists, means that the joint performance is greater than the solo performance of either i or j, i.e. $\lambda_{aa'} > \max\{\lambda_a, \lambda_{a'}\}$.

Finally, the stochastic term ξ_{ij} in equation (7) captures other unobserved factors that affect the final performance, and it is assumed to be normally distributed as $N\left(0,\sigma_\xi^2\right)$. Participants know the distribution but not the exact value of ξ_{ij} when making the team formation decisions.

The collection of performances of all teams under team structure $\mathcal{M}$ is $Y_{\mathcal{M}}$. The expected

[11] The benefits and costs of collaborations cannot be separately identified in our data, and as such $\lambda_{aa'}$ captures the net benefit in a reduced-form way.

monetary and non-monetary rewards of team $\langle i,j\rangle$, as equation (1) shows, depend on the rank of $Y_{\langle i,j\rangle}$ in $Y_{\mathcal{M}}$, i.e., $Z_{\langle i,j\rangle}(Y_{\mathcal{M}})$. Let $Pr\left(A_j = a|\langle i,j\rangle, R_j, R_i, \mathcal{R}\right)$ be the probability that the true ability of the target participant j is a, conditional on i and j teaming up, the target participant's signal R_j, the focal participant's signal R_i, and the collection of signals of all participants in the competition $\mathcal{R}$. This conditional probability also represents the updated belief of participant i regarding j's ability conditional on j agreeing to team up with i, which differs from the prior belief of j's ability, denoted by $Pr\left(A_j = a|R_j\right)$, that depends only on R_j. We will specify such conditional probability or the updated belief in the next sub-section.

Assume that the top P teams in the competition will receive monetary prizes, denoted by $Prize_p$ for the p^{th} place. The expected monetary reward for the focal participant can be specified as:

$$\begin{aligned} &E\left[Money\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right)\right)|A_i, R_i, R_j, \mathcal{R}\right] = \sum_{p=1}^{P}\left[Prize_p \times Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right) = p\right)\right] \\ &= \sum_{p=1}^{P}\left[Prize_p \times \sum_{a\in A} Pr\left(A_j = a|\langle i,j\rangle, R_j, R_i, \mathcal{R}\right) \times Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}|A_i, a, \mathcal{M}\right) = p\right)\right]. \end{aligned} \tag{8}$$

In the above equation, the probability that the rank of team $\langle i,j\rangle$ is at the p^{th} place is denoted on the right side in the first line. This probability is the sum of the conditional probability that A_j is equal to a specific level a multiplied by the probability that, given A_i and a as the true abilities of the two team members and team structure $\mathcal{M}$, the rank of the team is at the p^{th} place. This is expressed mathematically in the second line of the equation.

Similarly, we specify the expected non-monetary reward (i.e., Kaggle points) for team $\langle i,j\rangle$ as:

$$\begin{aligned} &E\left[Point\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right)\right)|A_i, R_i, R_j, \mathcal{R}\right] = \sum_{p=1}^{M}\left[Point_p \times Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right) = p\right)\right] \\ &= \sum_{p=1}^{M}\left[Point_p \times \sum_{a\in A} Pr\left(A_j = a|\langle i,j\rangle, R_j, R_i, \mathcal{R}\right) \times Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}|A_i, a, \mathcal{M}\right) = p\right)\right]. \end{aligned} \tag{9}$$

Note that the first summation on the right side of the equation is up to M, the total number of teams. This is because under Kaggle's policy, every team will receive a certain number of points.

The challenge of evaluating the expected monetary and non-monetary rewards is to compute the probability of the order, $Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}|A_i, a, \mathcal{M}\right) = p\right)$ in equations (8) and (9). The computation is complicated because it involves a rank order distribution. We use the asymp-

totic normality of the order statistic distribution to approximate the distribution of the performance of the p^{th}-place team. The asymptotic distribution mimics the actual probability very well when the number of participants is large in the competition. Utilizing this distribution function, we then use numerical method to compute $Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}|A_i, a, \mathcal{M}\right) = p\right)$. Details are in Appendix A.[12]

4.3 Updated Beliefs and the Market-Clearing Condition

Given signal R_j for participant j's ability, participant i's prior belief regarding j's ability is $Pr\left(A_j = a|R_j\right)$. Suppose j agrees to collaborate with i and let her take γ^M share of the monetary reward and transfer τ. We assume that i will update her belief with this new information, using the Bayes rule as follows:

$$Pr\left(A_j = a|\langle j,i\rangle, R_j, R_i, \mathcal{R}; 1-\gamma^M, -\tau\right) = \frac{Pr(\langle j,i\rangle|A_j, R_j, R_i; 1-\gamma^M, -\tau) \times Pr\left(A_j = a|R_j\right)}{\sum_{a'\in A} Pr(\langle j,i\rangle|a', R_j, R_i; 1-\gamma^M, -\tau) \times Pr\left(A_j = a'|R_j\right)}, \tag{10}$$

where $Pr(\langle j,i\rangle|A_j, R_j, R_i; 1-\gamma^M, -\tau)$ is j's choice probability given that her true ability is A_j, which is defined in equation (4). Note that, since γ^M and τ are what i takes from the team, j will receive $1-\gamma^M$ share of the monetary reward and $-\tau$ as transfer. Also, equation (10) implies rational expectation in the updated belief because it is based on j's optimal choice. Similarly, j will update her belief on i as follows:

$$Pr\left(A_i = a|\langle i,j\rangle, R_i, R_j, \mathcal{R}; \gamma^M, \tau\right) = \frac{Pr(\langle i,j\rangle|A_i, R_i, R_j; \gamma^M, \tau) \times Pr\left(A_i = a|R_i\right)}{\sum_{a'\in A} Pr(\langle i,j\rangle|a', R_i, R_j; \gamma^M, \tau) \times Pr\left(A_i = a'|R_i\right)}, \tag{11}$$

where $Pr(\langle i,j\rangle|A_i, R_i, R_j; \gamma^M, \tau)$ is also i's optimal choice probability given that her true ability is A_i. Combining the updated beliefs for both i and j in equations (10) and (11), we capture the following equilibrium effect: i's updated belief on j is based on j's optimal choice, while j's updated belief on i is also based on i's optimal choice; so both i and j's beliefs are consistent with each individual's optimal choice in the equilibrium.

When the market is at equilibrium, the number of participants with signal $r \in R$ who wants to match with participants with signal $r' \in R$ is equal to the other way around. Thus, γ^M and τ have to satisfy this market-clearing condition. However, researchers ob-

[12]In other empirical settings the payoffs for individual or firm collaborations may depend on the performance instead of the ranking. This will make the computation of the payoffs much easier without relying on order statistics as in rank-order tournaments. For example, when firms compete for market share, the payoff can be approximated by a multinomial logit market share function which is a function of the performances of the focal collaboration and other collaborations. In such case the payoff can be evaluated in an analytical way.

serve neither γ^M nor τ; as such, the two cannot be separately identified without additional data. We normalize γ^M to be $1/2$, and focus on solving for the market-clearing τ. This normalization does not affect the results because it can be derived from re-parameterization: assuming that the true sharing rule is $\widetilde{\gamma}^M \neq 1/2$ and the true transfer is $\widetilde{\tau}$, one can simply set $\gamma^M = 1/2$, and re-specify τ as $\widetilde{\tau}$ plus $\left(\widetilde{\gamma}^M - 1/2\right)$ multiplied by the expected monetary reward (preference for money θ^M is normalized to 1 in our empirical application), i.e., $\tau = \widetilde{\tau} + \left(\widetilde{\gamma}^M - 1/2\right) E[Money]$. The choice probability will remain unchanged. Under this normalization, the transfer τ captures the monetary value of the workload that i transfers to j relative to sharing the workload equally, plus the additional monetary reward that j pays to i relative to sharing the reward equally.

With the normalization, let $Pr(\langle r, r'\rangle|\mathcal{R}, \tau)$ be the probability that a participant with signal r chooses to collaborate with another participant with signal r', conditional on the collection of all participants' signals $\mathcal{R}$ and transfer τ. The probability can be derived as

$$Pr(\langle r, r'\rangle|\mathcal{R}, \tau) = \sum_{a \in A} Pr(\langle r, r'\rangle|a, r, r', \mathcal{R}; \gamma^M = 1/2, \tau) \times Pr(A = a|r), \tag{12}$$

where $Pr(\langle r, r'\rangle|a, r, r', \mathcal{R}; \gamma^M = 1/2, \tau)$ is the optimal choice probability on the right-hand side of equation (10) with the values of $\gamma^M = 1/2$ and τ. The market-clearing condition states that the transfer from the participant with signal r' to the participant with signal r, represented by $\tau(r, r')$, has to satisfy the following equality:

$$Pr(\langle r, r'\rangle|\mathcal{R}, \tau(r, r')) \times Pr_R(r) = Pr\left(\langle r', r\rangle|\mathcal{R}, -\tau\left(r, r'\right)\right) \times Pr_R(r'), \tag{13}$$

where $Pr_R(r)$ and $Pr_R(r')$ represent the proportions of participants with signals r and r', respectively.

Substitute equation (12) into (13), and further plug equation (3) into the equation, then apply logarithmic transformation and move terms, and we can derive that:

$$\begin{aligned}\tau(r, r') =& \frac{\mu}{2} \cdot \Bigg[\ln Pr_R(r') + \ln\Big(\sum_a \frac{\exp\left(V(a, r', r, \mathcal{R}; \gamma^M = 1/2, \tau(r, r'))/\mu + \tau(r, r')/\mu\right)}{\sum_{\tilde{r} \in (R \cup \varnothing)} \exp\left(V(a, r', \tilde{r}, \mathcal{R}; \gamma^M = 1/2, -\tau(r, \tilde{r}))/\mu\right)} \times Pr(A = a|r')\Big) \\ & - \ln Pr_R(r) - \ln\Big(\sum_a \frac{\exp\left(V(a, r, r', \mathcal{R}; \gamma^M = 1/2, \tau(r, r'))/\mu - \tau(r, r')/\mu\right)}{\sum_{\tilde{r} \in (R \cup \varnothing)} \exp\left(V(a, r, \tilde{r}, \mathcal{R}; \gamma^M = 1/2, \tau(r, \tilde{r}))/\mu\right)} \times Pr(A = a|r)\Big) \Bigg],\end{aligned} \tag{14}$$

where $V(a, r', r, \mathcal{R}; \gamma^M = 1/2, \tau(r, r'))$ is the expected payoff conditional on the values of $\gamma^M = 1/2$ and $\tau(r, r')$. This expression helps us to prove the existence of the market equilibrium (in the next sub-section). Finally, we provide a summary of all variables and

parameters in the model in Table 5.

Table 5: Model Variables and Parameters

Variables	Explanation
i, j	individual participant
$\mathcal{A}$	collection of abilities
$A_i = a$	i's ability is a
$\mathcal{R}$	collection/distribution of tiers
$R_i = r$	i's tier is r
$\langle i, j \rangle$	multi-member team
$\langle i, \emptyset \rangle$	single-member team
$\mathcal{M}$	the set of formed teams
p	rank in competition
$Money_p$	money reward for rank p
$Point_p$	point reward for rank p
$Y_{\langle i,j \rangle}(a, a')$	$\langle i, j \rangle$'s performance with abilities a, a'
ξ_{ij}	unobserved team performance factor
$Y_{\mathcal{M}}$	collection of performances of all teams
$Z_{\langle i,j \rangle}(Y_{\mathcal{M}})$	rank of team $\langle i, j \rangle$ under $Y_{\mathcal{M}}$
$Pr(Z_{\langle i,j \rangle}(Y_{\mathcal{M}}) = p)$	probability of team $\langle i, j \rangle$'s rank equals p under $Y_{\mathcal{M}}$
$Money(Z_{\langle i,j \rangle}(Y_{\mathcal{M}}))$	money reward of team $\langle i, j \rangle$ under $Y_{\mathcal{M}}$
$Point(Z_{\langle i,j \rangle}(Y_{\mathcal{M}}))$	point reward of team $\langle i, j \rangle$ under $Y_{\mathcal{M}}$
γ^M	sharing rule of monetary reward
γ^P	sharing rule of point reward
$\tau(R_i, R_j)$	workload transfer from tier R_i to R_j
$U(A_i, R_i, R_j, \mathcal{R})$	utility for i of A_i,R_j when teaming with j of R_j under $\mathcal{R}$
ε_{i,R_j}	unobserved utility factor between i & tier R_j
$Pr(\langle i, j \rangle \vert A_i, R_i, R_j, \mathcal{R}; \gamma^M, \tau)$	choice probability for i of A_i,R_i teaming with j of R_j under $\mathcal{R}$, conditional on endogenous γ^M, τ
$Pr(A_j \vert R_j)$	prior belief of tier R_j's ability
$Pr(A_j = a \vert \langle i, j \rangle, R_j, R_i, \mathcal{R}; 1 - \gamma^M, -\tau)$	i's posterior belief on j's ability equaling a if j of tier R_j agrees to form team with i of R_i under $\mathcal{R}$, conditional on j receiving $1-\gamma^M, -\tau$ in share of money and workload transfer
$\boldsymbol{Pr(\mathcal{M})}$	collection of choice probabilities of all tiers and abilities
Parameters	**Explanation**
θ_i^M, θ_i^P	marginal utility for money and points
$\lambda_{aa'}$	team performance
$\alpha(R_i, R_j)$	net benefit in team between tier R_i,R_j
$\boldsymbol{\tau}$	collection of transfers
$\gamma^M(R_i, R_j, \mathcal{R})$	tier R_i's share of money with R_j under $\mathcal{R}$
μ	scaling parameter of ε under extreme type I distribution
δ_ξ	standard deviation of ξ under normal distribution

4.4 Market Equilibrium

The matching game in our model is characterized by the preference parameters $\left\{\theta_i^M, \theta_i^P, \alpha\left(R_i, R_j\right)\right\}$ for every participant (see equations (1) and (2)), the monetary and non-monetary rewards $\{Prize_p, Point_p\}$ for every rank (see equations (8) and (9)), and how the non-monetary rewards are split, i.e., γ^P in equation (1). The market is at equilibrium when the market-clearing condition in equation (13) is satisfied for every (r, r') pair. In addition, the probability that a participant with ability and signal (A_i, R_i) matches with another with signal r has to be the participant's optimal choice. That is, equation (3) has to be satisfied when

$\gamma^M = 1/2$ and $\tau = \tau(r, r')$. The market equilibrium is represented by the choice probabilities in equations (10) and (11) and transfer τ.

Let $\boldsymbol{Pr(\mathcal{M})}$ be a vector with $(R \times A) \times (R + 1)$ rows that represents the collection of the choice probabilities of all ability and signal types (including single-member team choice), and $\boldsymbol{\tau}$ be a vector with $(R^2 + 1)$ rows that represents the collection of transfers from one to another signal type within a team (the transfer in a single-member team is fixed to zero). We can combine equations (3) and (14) into a system of equations $\boldsymbol{\mathcal{H} : (Pr(\mathcal{M}), \tau) \to (Pr(\mathcal{M}), \tau)}$:

$$\begin{cases} \boldsymbol{Pr(\mathcal{M}) = h_1(Pr(\mathcal{M}), \tau)}, \\ \boldsymbol{\tau = h_2(Pr(\mathcal{M}), \tau)}. \end{cases} \tag{15}$$

Definition 1 *A **Competitive Collaborative Market Equilibrium** $\boldsymbol{(Pr(\mathcal{M})^*, \tau^*)}$ is the solution of the equation system $\boldsymbol{\mathcal{H}}$.*

Proposition 1 *For each competition characterized by $\left\{\theta_i^M, \theta_i^P, \alpha(R_i, R_j)\right\}$ for every participant, $\{Prize_p, Point_p\}$ for every rank, and γ^P for every competition, market equilibrium defined as the solution of the equation system $\boldsymbol{\mathcal{H}}$ in equation (15) exists.*

The proof is in Appendix B.

The equilibrium concept is based on the matching model with transferable utilities in Becker (1973) and the empirical matching model in Choo and Siow (2006), both applying to the marriage market. One important guarantee of the existence of the equilibrium is the large market assumption, i.e., there are enough agents for each type such that one agent could always find her most favorable type of agent if the agent is willing to pay the equilibrium transfer. Our empirical setting satisfies the assumption, as there is a large number of participants (hundreds or even thousands) for each public tier in each competition. As discussed above, our study differs from these earlier papers, as we allow for incomplete information in the empirical model. Agents form rational beliefs due to incomplete information. The updated beliefs in equations (10) and (11) illustrate that both i and j form rational expectations when forming a team.

Note that our model does not utilize the pairwise stability conditions in the model estimation as in Fox (2008) and Wu (2015). Instead, we use the maximum likelihood estimation while imposing marketing-clearing conditions for the transfers $\boldsymbol{\tau}$, which is similar to Chan et al. (2015). However, under the large market assumption, the matching probability functions in equations (3) and (4) imply that neither i nor j wants to switch to team up with other agents of different tiers conditional on the beliefs in the equilibrium. This indicates that the equilibrium is not blocked by the deviation of any two tiers conditional on on-path beliefs. If a deviation conditional on rational off-path beliefs is also consistent with the

market-clearing conditions, it means that we will have multiple equilibria. While this could be possible, in Appendix D, we use numerical simulations to empirically show that the equilibrium is unique in the range of reasonable parameters. Therefore, the pairwise stability condition is implicitly satisfied.

5 Model Estimation

In this section, we discuss how we operationalize the model estimation in the empirical context and provide model-identification intuitions.

In the model, each participant is characterized by her true ability and a signal about her ability. We discretize participant ability into three levels, i.e., $A = \{Low, Medium, High\}$. We use the public Kaggle tier status to proxy the noisy signal. That is, $R = \{Novice, Contributor, Expert, M$ As the most highlighted information on the participant's default public profile page, tier status is the simplest source for players to differentiate the abilities of potential teammates. Other signals, such as a participant's accumulated Kaggle points, are also accessible on the detailed profile page; yet they are highly correlated with the tier status and the differences between the points are more difficult to interpret. Our model estimation is also much simplified with a finite discrete measure of participant ability signal.

There are three sets of model parameters: utility parameters that are linked with the observed team formation decisions, parameters for the mapping between public tier information and private ability, and parameters on team abilities driving the observed performances. For the utility parameters, we normalize the marginal utility of the monetary rewards θ_i^M to 1, and allow the preference for non-monetary point rewards θ_i^P to vary based on the participant's tier status, i.e., $\theta^P = \{\theta^P_{Novice}, \theta^P_{Contributor}, \theta^P_{Expert}, \theta^P_{Master}\}$. Such heterogeneity captures the fact that a participant's valuation of Kaggle points may differ across tiers. We also allow θ^P to change before and after Kaggle adjusted its point-allocation rule. This reflects the fact that the points a participant can earn from a competition are significantly different after the policy change.

We allow the benefit from collaboration, i.e., $\alpha(R_i, R_j)$ in equation (1), to differ across each unique combination of participants' public tier status (R_i, R_j), but we assume $\alpha(R_i, R_j) = \alpha(R_j, R_i)$. The symmetric benefit assumption is solely a normalization for identification purposes. To understand this, suppose i receives more benefits from forming a team with j, i.e., $\alpha(R_i, R_j) > \alpha(R_j, R_i)$, then i and j could negotiate the transfer $\tau(R_i, R_j)$ such that i will agree on a larger transfer to j; otherwise, more participants of type R_i will want to collaborate with type R_j and the market will not be at equilibrium. In other words, under the market clearing assumption, the difference between $\alpha(R_i, R_j)$ and $\alpha(R_j, R_i)$ can not be

separately identified from the transfer $\tau(R_i, R_j)$ and thus the assumption is necessary. Consequently, there are 10 α's to be estimated with four public tiers. Under this normalization, one should interpret $\alpha(R_i, R_j) + \tau(R_i, R_j)$ as the negotiated total benefit that i obtains from working with j, relative to competing solo. In Online Appendix C, we use numerical simulations to further illustrate this argument. Regarding the transfer τ's, we do not put further restrictions on the values but allow them to vary flexibly across competitions and team types. Indeed, they are not global model parameters to be estimated; instead, they are determined by the market equilibrium conditional on the assumed model parameters (equation 14) in the inner level of the model estimation (details are below). Given any model parameters, the values of τ's are bounded; otherwise the market will not be at the equilibrium.

Since we do not observe the actual ability, the mapping function $Pr(A_i = a|R_i)$ in equation (10) also needs to be estimated from the data. This probability is specific to every combination of public tier and private ability and sums up to one for each public type; therefore, there are $4 \times 3 - 4 = 8$ parameters with four tiers and three ability levels. As a reduced-form way of capturing how competitions with various prize levels may attract different pools of talent to participate, we also allow the probabilities to differ for competitions with small and large monetary rewards.[13]

For the team performance function specified in equation (7), we estimate the team ability λ's for each unique combination of actual player ability types (a, a'), including those for competing solo. There are 9 λ's to be estimated with three ability levels. The difference between the λ's for a team ($\lambda_{aa'}$) and its individual members ($\lambda_{a\emptyset}, \lambda_{\emptyset a'}$) reveals the potential benefits and challenges of team work in performance. We estimate individual team ability parameters for each unique combination of player types instead of imposing parametric restriction, because team ability may not be a simple addition or linear combination of individual member abilities; rather, it could depend on with whom a participant is collaborating with. It is even possible that the team ability could be worse than that of its members when competing solo, e.g., if communication efforts take up much working time. Finally, we also estimate the scale parameter μ in the utility function and the variance parameter δ_ξ^2 in the performance function.

5.1 Maximum Likelihood under Equilibrium Constraints

The outcomes of the matching game we observe from the data include team formation $\mathcal{M}$ and performance $Y_{\langle i,j \rangle}$ of every team. Given a transfer τ, the probability of i collaborating

[13]Since competitions with high monetary awards are usually sponsored by bigger businesses and the tasks are more challenging, there may be other non-monetary incentives that attract talents to participate in these competitions.

with j is:

$$\begin{aligned}\mathcal{L}(\langle i,j\rangle|R_i,R_j,\mathcal{R},\tau) =&(\sum_{a\in A} Pr(\langle i,j\rangle|a,R_i,R_j,\mathcal{R};\gamma^M=1/2,\tau)\times Pr(A_i=a|R_i))\times\\ &(\sum_{a'\in A} Pr(\langle j,i\rangle|a',R_j,R_i,\mathcal{R};\gamma^M=1/2,-\tau)\times Pr(A_j=a'|R_j)).\end{aligned} \tag{16}$$

The equation indicates that the team will be formed only if it is optimal for both participants.[14]

The likelihood of observing team performance $Y_{\langle i,j\rangle}$ conditional on a team $\langle i,j\rangle$ being formed is:

$$\begin{aligned}\mathcal{L}(Y_{\langle i,j\rangle}|R_i,R_j,\mathcal{R},\tau) = \sum_{a\in A}\sum_{a'\in A} &Pr(A_i=a|R_i,R_j,\mathcal{R};\gamma^M=1/2,\tau)\times\\ &Pr(A_j=a'|R_j,R_i,\mathcal{R};\gamma^M=1/2,-\tau)\times\\ &\phi(Y_{\langle i,j\rangle},\lambda(a,a'),\delta_\xi^2),\end{aligned} \tag{17}$$

where $\phi(y,\lambda(a,a'),\delta_\xi^2)$ is the normal probability density function of the performance when the teammates' abilities are a and a', and the conditional probability function comes from equation (10).

The likelihood function we use in model estimation is the sum of the log-likelihoods of the observed team formation and performance in every competition in the data. That is,

$$l(\Theta) = \sum_g \sum_{\langle i,j\rangle\in\mathcal{M}_g} [l(\langle i,j\rangle|R_i,R_j,\mathcal{R},\tau) + l(Y_{\langle i,j\rangle}|R_i,R_j,\mathcal{R},\tau)], \tag{18}$$

where Θ denotes the set of model parameters, subscript g a competition and $\mathcal{M}_g$ the collection of all teams in the competition. In addition, $l(\langle i,j\rangle|R_i,R_j,\mathcal{R},\tau)$ and $l(Y_{\langle i,j\rangle}|R_i,R_j,\mathcal{R},\tau)$ are the logarithm transformations of the likelihoods in equations (16) and (17), respectively.

The challenge of evaluating equation (18) is two-fold. First, the transfer vector $\boldsymbol{\tau}$ is unobserved to researchers; it has to be recovered from the market-clearing conditions in every competition. Second, the payoff from working in a team depends on how other teams in the competition are formed, i.e., the matching probability of a team is a function of the matching probabilities of other teams in the competition, as the first line of function $\boldsymbol{h_1}$ in equation (15) suggests. Because of these two complications, the likelihood function cannot be evaluated analytically.

[14] For a single team $\langle i,\varnothing\rangle$, the second line on the right side is fixed at 1.

We propose a two-level estimation procedure to tackle the challenge. In the inner level, conditional on trial parameters Θ we search for the matching probabilities and transfers $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ for every competition such that $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^* = \boldsymbol{h_1}(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ and $\boldsymbol{\tau}^* = \boldsymbol{h_2}(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$. That is, we find $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ that satisfy the equilibrium constraints. In the outer level, we search for Θ that maximizes the likelihood function in equation (18). The algorithm proceeds as follows:

1. Start with initial parameter vector Θ^0,
   (a) With initial values $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^{\boldsymbol{0}}, \boldsymbol{\tau}^{\boldsymbol{0}})$, calculate expected payoffs using numerical methods;
   (b) Calculate $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})', \boldsymbol{\tau}') = \boldsymbol{\mathcal{H}}(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^{\boldsymbol{0}}, \boldsymbol{\tau}^{\boldsymbol{0}})$;
   (c) Replace $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^{\boldsymbol{0}}, \boldsymbol{\tau}^{\boldsymbol{0}})$ by $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})', \boldsymbol{\tau}')$. Repeat the above procedure (a)-(b) until $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}}), \boldsymbol{\tau})$ converge. They represent the market equilibrium under model parameters Θ^0;
   (d) Calculate the likelihood function value in equation (18) under parameters Θ^0.
2. Search for Θ and evaluate the likelihood function value with the above process and find Θ such that the likelihood function is maximized.

Note that Proposition 1 proves the existence but not the uniqueness of the equilibrium. Potential multiple equilibria are a concern when we estimate the model and conduct counterfactuals. During model estimation, we test whether this is an issue by varying the starting values of $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^{\boldsymbol{0}}, \boldsymbol{\tau}^{\boldsymbol{0}})$ in the inner level. We find that they always converge to the same $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$, suggesting that the equilibrium is unique in our empirical application. We provide details in Online Appendix D.

5.2 Identification

Given that there is no analytical solution to the equilibrium, we provide an informal discussion on the intuitions behind model identification arguments in this sub-section.

The identification of the parameters of the team performance function (λ, δ_ξ^2), and the probability of true ability A conditional on public signal R, $Pr(A|R)$, comes from how team scores vary across different combinations of R_i and R_j, as equations (17) and (10) suggest. In the likelihood, $\mathcal{L}(Y_{\langle i,j\rangle})$ can be treated as a latent class regression, with $Pr(A|R)$ representing the size of the latent classes.

The proportions of teams with members from different public tiers identify the net benefits of team formation, α. Conditional on the expected $Y_{\langle i,j\rangle}$ and thus the monetary and

non-monetary rewards, the larger the proportion of teams formed by tiers R_i and R_j, the larger the value of $\alpha(R_i, R_j)$, relative to $\alpha(R_i, \varnothing)$ and $\alpha(R_j, \varnothing)$ that are normalized to zero. The identification of preferences for Kaggle points, θ^P, comes from the proportions of teams formed across types of tiers and across competitions. For Kaggle competitions, points allocated to a team increase by both the ranking of the team performance and the number of participants. As an illustration, assuming that Master-Master teams on average perform better than Masters competing solo, or Masters teaming with other tiers. Suppose that for competitions with a large number of participants (thus high-ranked teams will obtain more Kaggle points), Masters are more likely to form teams with other Masters. In contrast, for other competitions when there are not many Kaggle points to be allocated, Masters are more likely to form teams with participants from lower tiers (and thus enjoy the transfer of non-points benefits from the teammate). This suggests that θ^P is high for Masters.

Finally, conditional on monetary prizes and Kaggle points, the variation in the proportions of teams formed by different types of participant tiers identifies the scale parameter μ. Suppose, for example, as the monetary prize increases across competitions, the proportions only vary slightly. This lack of variation in team composition will imply a high value of μ.

6 Results

In this section, we report the estimation results. The estimated marginal utilities for Kaggle points (θ^P) are presented in Table 6. As the marginal utility for the monetary prize is normalized to one, the estimates in the table represent how much a participant would value one Kaggle point. Because of the change in the point-allocation policy we described in Section 3, the value of Kaggle points may adjust correspondingly; therefore, we separately estimate the marginal utilities before and after the policy change.

Table 6: Parameter Estimates of Preferences for Kaggle Points

Parameter	Public Tier			
	Novice	Contributor	Expert	Master
θ^P: before policy change	0.36 (0.11)	1.53 (0.18)	0.82 (0.05)	0.41 (0.12)
θ^P: after policy change	2.02 (0.04)	8.11 (0.06)	4.98 (0.26)	3.77 (0.02)

Note: Numbers represent the point estimates; numbers in parentheses are the standard errors of the point estimates.

There is an inverted-U shaped relationship between the marginal utility for points and participant tiers. The marginal utility increases as participants progress from Novice to Contributor, then decreases as they further progress to Expert and Master. The result is

probably due to the way Kaggle determines tier status. Novices are new entrants to the platform who have not participated in any competition. In the data, more than 80% of Novices participated in only one competition. As the Kaggle points from one competition are unlikely to benefit these individuals in the future, their average marginal utility for points may be low. The remaining 20% progress to the Contributor tier when they participate again. They are those who self-select to continue to compete; therefore, they may have a much higher marginal utility for points. The decreasing marginal utility for points from Contributor to Master probably reflects the fact that the value of gaining additional points is lower, as participants have accumulated more points. Still, the marginal utilities of Experts and Masters are significantly positive, suggesting the value of non-monetary benefits of accumulating more points so that they will rank high on the Kaggle leaderboard (see the discussion in the data section). Interestingly, marginal utilities for Kaggle points increase substantially after the policy change. One possible reason is that, as more and more participants are attracted to join competitions in later periods, it has become more difficult to win Kaggle points. The result also explains why collaborations have dropped after the policy change in the data: as Kaggle points are valued more, participants are more reluctant to form teams lest they have to split points with the others.

The results suggest that Kaggle points are quite valuable for participants. Using the estimates multiplied by the average number of points a participant wins in competitions, we see that the average value of Kaggle points earned in each competition is $7,600 before and $4,900 after the policy change. As a benchmark, the expected monetary reward for an average participant is just $39. The comparison suggests that for most participants, non-monetary payoff dominates monetary payoff, consistent with our quote in the data section from Gilberto Titericz, a top Kaggle player, regarding the benefits of Kaggle points. To better understand where the benefits come from, we further examined the related discussions on Quora.com, a popular question-and-answer website, and found various answers suggesting that being placed high in the tier system or on the leaderboard helps enrich one's resume and increase the chance of getting hired with a high salary offer.[15] As a high ranking on Kaggle requires consistently gaining high points across competitions, we believe the high marginal utility of Kaggle points reflects the improved value of future career opportunities.

Table 7 reports the estimated net benefits ($\alpha(R, R')$) from forming teams on top of earning monetary prizes and Kaggle points. All the estimates are significantly positive, implying that

[15]For example, one answer for the question "Can someone make a living from solving problems on sites like Kaggle" (https://www.quora.com/Can-someone-make-a-living-from-solving-problems-on-sites-like-Kaggle) says that "I see more people benefiting from Kaggle by putting their ranks in resumes. Having a decent ranking certainly sets you apart from the crowd. I won't be surprised if you end up with a 30% to 50% jump in salaries with a decent ranking."

in general, the benefits of forming teams outweigh the costs. These benefits are also higher than the expected monetary reward for an average participant, suggesting that, in addition to Kaggle points, the non-monetary benefits from peer collaborations are important. Since collaboration could bring in economy of scale and division of labor, these benefits could come from the time and effort savings. For instance, when two novices work together, they can split the workload, so the $151 benefit reflects the saving of workload relative to working alone (in this case, the transfer τ is 0). Furthermore, the net benefits from teaming with individuals in high tiers are higher than with those in low tiers. For example, the benefit for an Expert working with a Master is $407, more than double that of working with a Novice ($183). The additional benefit can come from the value of learning from the Master teammate.[16]

Table 7: Parameter Estimates of Preferences for Collaborations

Team Structure	Novice	Contributor	Expert	Master
Novice	151 (0.27)	162 (0.03)	183 (0.31)	220 (0.16)
Contributor		276 (0.05)	289 (0.05)	310 (0.04)
Expert			378 (0.32)	407 (0.04)
Master				552 (0.17)

Note: Numbers represent the point estimates; numbers in parentheses are the standard errors of the point estimates.

Table 8 reports the estimated probability of a participant's ability level conditional on her tier status (i.e., $Pr(A|R)$). We group the competitions in our data into low- and high-prize types, using the average prize amount across competitions as the criterion. There are 77 low-prize and 25 high-prize competitions, with the average prize amount about $9,000 and $76,000, respectively. A low-prize competition attracts about 400 participants whereas a high-prize one attracts about 1,200 participants. We separately estimate the conditional probabilities for these two types of competitions. Results show that Kaggle's tier system is in general consistent with participants' ability. For example, the proportion of high-ability individuals increases from 33-38% for Novices to 90-94% for Masters. However, the variation in abilities within each tier is also substantial, indicating that tiers are a noisy signal. This is especially the case for Novices, as the proportions of individuals with low and high-ability are both large. We further calculate the distribution of true abilities in the pool of participants, by summing up the products of the conditional probability and the size of each tier, across

[16]We will show below that there is a positive transfer of value from the Expert to the Master ($1,449) in an Expert-Master team. This implies that the total net benefit for the Expert to collaborate with the Master is even larger than $407; otherwise, the Expert will not agree to the transfer. In contrast, the benefit for the Master working with the Expert is much smaller.

all tiers. We report these in the rows of "Total". The sizes of low-ability and medium-ability groups are about the same, and both are about half the size of high-ability participants, indicating that Kaggle's competitions are attractive to top talents. In comparison with small-prize competitions, the proportion of individuals with high-ability is larger in high-prize competitions, while the proportion of low- and medium-ability individuals is smaller. Competitions with high monetary rewards are usually hosted by big-name companies with more challenging and impactful problems, thus they attract more talents.

Table 8: Parameter Estimates for Conditional Probability

Public Type	Private Type		
	Low Ability	Medium Ability	High Ability
Small Reward Games			
Novice	0.41 (0.04)	0.26 (0.09)	0.33
Contributor	0.18 (0.07)	0.47 (0.12)	0.35
Expert	0.09 (0.14)	0.20 (0.13)	0.71
Master	0.03 (0.18)	0.07 (0.16)	0.90
Total	0.29	0.27	0.44
Large Reward Games			
Novice	0.37 (0.09)	0.25 (0.11)	0.38
Contributor	0.24 (0.11)	0.28 (0.07)	0.48
Expert	0.06 (0.12)	0.12 (0.14)	0.82
Master	0.03 (0.21)	0.03 (0.11)	0.94
Total	0.27	0.22	0.51

Note: Numbers represent the point estimates; numbers in parentheses are the standard errors of the point estimates; numbers in rows of "Total" represent unconditional probability of each ability type.

Table 9 reports the estimated mean performance parameter (i.e., $\lambda(a, a')$) of each type of teams. Estimates in the last column of the table are the mean performance of single-member teams, which can be used as the benchmark against the performance of collaborations. There is a strong increase in performance from low to high-ability single-member teams. Comparing the left columns with the last column, one can see that collaborations clearly help improve the team performance. For example, the predicted performance of a low-low (high-high) combination is 0.52 (2.12), much higher than that when a low-ability (high-ability) participant works alone. However, there is a downside for high-ability participants: if they work alone, the predicted performance is higher than if they team up with medium- or low-ability individuals. This difference is probably because potential substandard work from a low-ability member can have a substantial impact on the whole performance of the team. These results imply that the lack of information regarding the true ability of other

participants can become a hurdle for high-ability individuals to form teams. Therefore, improving the informativeness of the tier system may create better opportunities for the collaborations among high-ability individuals, a result we will show in the next section.

Table 9: Parameter Estimates for Team Performance

Team Structure	Low Ability	Medium Ability	High Ability	Single
Low Ability	0.52 (0.02)	0.78 (0.02)	1.62 (0.09)	-2.21 (0.34)
Medium Ability		0.89 (0.12)	1.77 (0.35)	-1.07 (0.14)
High Ability			2.12 (0.39)	1.82 (0.17)
σ_ξ^2	0.53(0.16)			
μ	1313(21.7)			

Note: Numbers represent the point estimates; numbers in parentheses are the standard errors of the point estimates.

Finally, the estimated variance of the team performance (δ_{xi}^2) is 0.53, and the scale parameter in participants' utility function (μ) is 1,313. Both are relatively large compared with the predicted team performance and average payoff levels, respectively. This indicates large randomness in both team formation and performance.

Using model estimates, we could recovered the equilibrium transfers τ within each competition. The average transfer from one tier to another across competitions is reported in Table 10. Lower tiers have to pay a positive transfer to higher tiers when forming teams, and the magnitude increases as the difference in tiers increases. For instance, to form a team with a Master, an Expert on average needs to pay \$1,449, while a Contributor needs to pay more than twice the amount at \$3,046. The high transfers reflect the high non-monetary benefits from forming teams with high-ability teammates. When teaming up with an Expert, for example, the average value from gaining more Kaggle points and other benefits (i.e. α) is \$9,073 for Contributors but only \$6,841 for an Expert. The Contributor, therefore, is willing to transfer a positive value to the Expert. The supply of the participants in different tiers is another reason for the amount of transfers. As shown in Table 2, there are more Contributors(18.7%) and Experts(11.5%) than Masters(9.8%). If the transfers are too low, there will be more Contributors and Experts who want to form teams with Masters and the market will not clear.

Our model also allows us to recover the choice probabilities of participants given the equilibrium transfers. Importantly, depending on participants' true abilities, their choice probabilities differ even when they belong to the same public tier. For instance, the probability of low-ability Novices teaming with another Novice is 42%, with a Master the probability is 1%, and there is a 43% probability that the Novice will work solo. The probabilities for a high-ability Novice are 6%, 4% and 88%, respectively. A high-ability Novice is more likely

Table 10: Average Transfers Between Participants

Paid by	To Teammate			
	Novice	Contributor	Expert	Master
Novice	0	2701	3477	4231
Contributor		0	1088	3046
Expert			0	1449
Master				0

Note: Transfer for participants within the same tier is 0. Positive values in the upper triangle mean that participants from a lower tier will pay positive transfer to participants in a higher tier.

to stay single, because the loss from splitting points with the teammate is potentially large. However, if she can team up with another high-ability teammate, she can have a high chance to achieve a top ranking and thus win both monetary and non-monetary rewards. Therefore, she is more likely to collaborate with a Master.

Panel (A) of Table 11 compares the average percentage of team types observed in the data with that predicted by our model. Overall, the predicted distribution of team types is highly consistent with the data pattern. The only collaboration that the model significantly under-predicts is Masters teaming with Masters. Panel (B) of Table 11 further compares the average scores across each team type in the data with those predicted by our model.[17] The numbers are again quite close to each other. For example, The model is able to replicate how collaborating with a teammate of the same or different tier, a Master is able to obtain an average score higher than competing solo. Overall, our model is able to predict well the team formation and performance as observed in the data.

7 Counterfactuals

In this section, we will use the estimation results to examine through counterfactual analyses how the incomplete information and competitive pressure impact team formation and performance. Specifically, we manipulate the informativeness of public tiers and the point allocation policy at Kaggle in separate counterfactuals. The results can help the platform to design better policies to further motivate collaborations, which could ultimately benefit participants and sponsoring businesses.

[17]Note that Table 9 reports the predicted performance based on the true abilities of two teammates. Predictions from Table 11 are based on the distribution of true abilities for each tier, and the distribution of team types based on the tiers of two teammates.

Table 11: Model Fit

Tier/Choice	Single	Team			
		Novice	Contributor	Expert	Master
(A) Team Types					
Novice	40.2% (41.6%)	15.8% (15.7%)	2.4% (3.0%)	1.1% (1.4%)	1.3% (1.4%)
Contributor	17.2% (14.7%)		1.0% (0.9%)	0.4% (0.4%)	0.2% (1.4%)
Expert	10.6% (9.5%)			0.1% (0.7%)	0.1% (0.7%)
Master	7.4% (7.6%)				2.2% (1.0%)
(B) Team Scores					
Novice	0.64 (0.68)	0.77 (0.80)	0.76 (0.79)	0.95 (0.87)	1.32 (1.21)
Contributor	0.61 (0.68)		0.83 (0.81)	0.95 (0.89)	1.07 (1.16)
Expert	0.82 (0.82)			1.13 (1.01)	1.19 (1.31)
Master	1.03 (1.11)				1.43 (1.44)

Note: Panel (A) represents percentage of team type; panel (B) represents average performance for team types; numbers represent values from data; numbers in parentheses represent predicted values from estimation.

7.1 Impacts of the Uncertainty Concerning a Teammate's Ability

Our estimation results (see Table 8) show that tiers are an imperfect signal of an individual's true ability. In the first counterfactual we explore the impacts of changing the tiers' informativeness on how participants form teams and on their performances.

We select one competition held with the original Kaggle point policy that offered a $50,000 reward and attracted 1,396 participants in the data for the analysis. The Novices, Contributors, Experts, and, Masters constitute 71%, 13%, 9%, and 7%, respectively. Our model recovers a 30%, 23%, and 47% split for the low-, medium-, and high-ability participants, respectively. We manipulate the informativeness of the tier system by changing the conditional probability of a participant's ability across different tiers and simulate the team formation and performance outcomes. We refer to the information structure from model estimates as the original-information scenario (see the three columns under "Original Information" in Table 12). We also create a no-information scenario with equal probabilities across tiers (see the three columns under "No Information") and a high-information scenario with a much stronger correlation between the tier and ability (see the three columns under "High Information"). Moving from left to right in Table 12, tier status becomes more informative and thus reduces the uncertainty about potential teammates' ability. For example, Novices become more likely to belong to the low-ability type (increase from 30% to 39%), while Experts and Masters are more likely to be the high-ability type (increase from 47% to 92% and 98%, respectively). In each scenario, we simulate the equilibrium outcomes by

solving for the fixed point in equation (15). The choice probabilities and transfers are also computed in each scenario.

Table 12: Different Information Scenarios ($Pr(A|R)$)

Public Type	No Information			Original Information			High Information		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
Novice	0.30	0.23	0.47	0.37	0.25	0.38	0.39	0.23	0.38
Contributor	0.30	0.23	0.47	0.24	0.28	0.48	0.14	0.48	0.38
Expert	0.30	0.23	0.47	0.06	0.12	0.82	0.06	0.02	0.92
Master	0.30	0.23	0.47	0.03	0.03	0.94	0.01	0.01	0.98

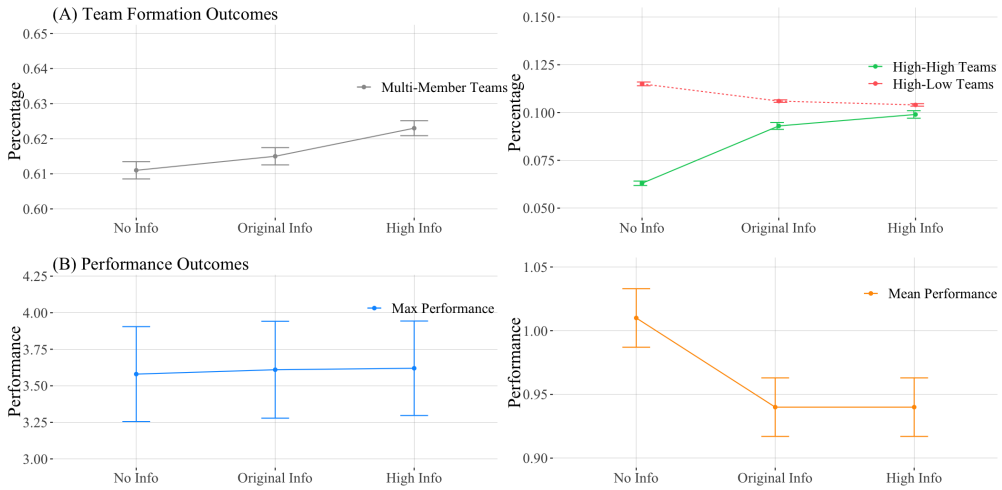
Figure 2 (A) graphically compares the team formation outcomes across the three scenarios. The percentage of multi-member teams increases from 61.1% under the no-information scenario to 61.5% under the original-information scenario, and further to 62.3% under the high-information scenario. This indicates that increasing the informativeness of tier status will enhance the collaboration of participants. The increase in collaboration mainly comes from more high-high collaborations, i.e. collaborations between two high-ability participants, while high-low collaborations will decrease (see the right diagram of Figure 2 (A)). The result suggests that more precise information regarding the ability of potential teammates facilitates the matching between high-ability participants and reduces "mis-matches" of abilities in the team formation.

Diagrams of Figure 2 (B) compare team performances in the three scenarios. The maximum performance of all the teams increases from 3.58 in the no-information scenario to 3.62 in the high-information scenario (see the left diagram). The result suggests that increasing the informativeness of tier status benefits sponsoring businesses, since for them the outcome of the winning team is most valuable and outcomes from other teams are unlikely to be adopted. The average performance across all teams however declines from the no-information to the high-information scenario (see the right diagram). The decline is due to the fact that low-ability individuals find it more difficult to team up with high-ability individuals when the tier status is informative.

Overall, the results suggest that uncertainty about the ability of teammates is an important factor that hinders collaborations between high-ability individuals. They would face a great chance in matching with low-ability individuals when public information is uninformative. If Kaggle focuses on delivering the best outcomes for sponsoring businesses, it should improve the informativeness of the tier system; however, doing so may reduce the collaboration opportunities for low-ability participants, limit their learning from more capable peers,

and lower the performance of an average participant[18].

Figure 2: Impact of Informativeness



Bars represent 90% confidence intervals estimated based on 100 rounds of simulations.

7.2 Impacts of Kaggle's Point Allocation Policy

Our estimation results suggest that players value the Kaggle points they earn in competitions. Thus, Kaggle's policy on point allocations substantially impacts the payoffs of participants, and because points are awarded based on ranks, it will also create competitive pressure on teams. In the second counterfactual analysis, we study how the competitive pressure impacts team formation and performance.

The point-allocation policy can be represented by the slope of the point function that determines how many points a team can earn at different performance rankings. A flatter curve means that points are allocated more evenly across teams, while a steeper curve puts more weight on the performance ranking, such that top-ranked teams receive significantly more points. A steeper curve also creates more competitive pressure because it gives teams extra motivation to stay on the top. We use the average points difference between two neighboring performance rankings as the measure of the slope, which is equivalent to the average absolute value of the gradients across all different ranks. Under Kaggle's current policy (see the green solid curve in Figure 3), the slope is 204, indicating that by moving one rank higher, a team will on average obtain 204 more Kaggle points. We then construct two counterfactual policies with different slopes. To make sure that the results are driven

[18]Kaggle recently introduced a multi-dimensional tier system (tiers based on different attributes), a direction that helps improve the informativeness of tiers. See `https://www.kaggle.com/progression/` for details.

only by the change in the slope of the allocation function, we fix the total points awarded to all teams in all scenarios to be 7 million, the same as what we observe in the data. We first create a counterfactual "flat-slope" scenario by setting the slope to 22 (see the dashed red curve in Figure 3), under which improving the performance ranking does not bring the team many Kaggle points. In another "steep-slope" scenario, the slope is 5,097. The dotted blue curve in Figure 3 shows that a large number of points are awarded to the top few teams, and teams ranked 50 or lower will gain very few points. We use the same competition as in Section 7.1 in this counterfactual.

Figure 3: Point Allocation Slopes

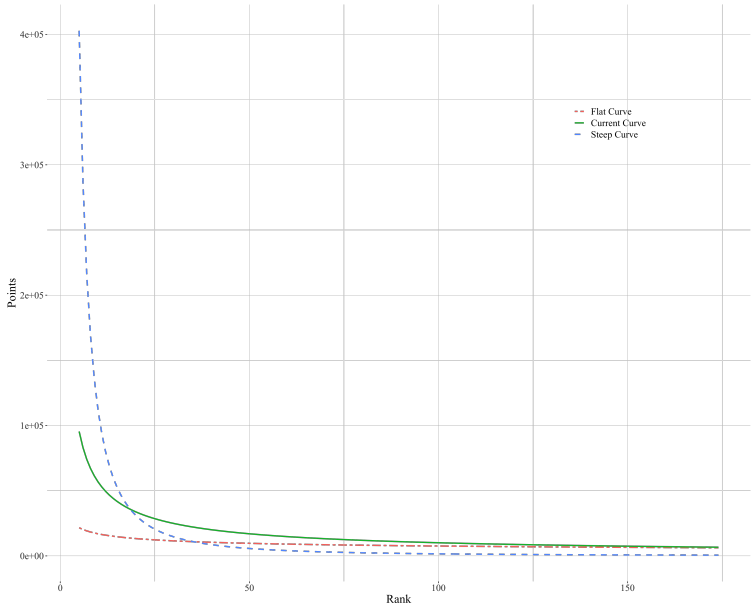
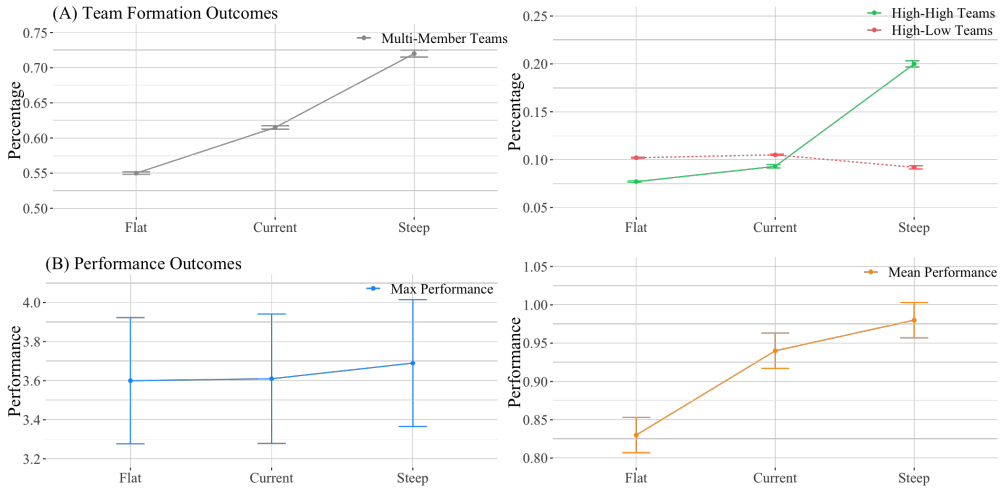


Figure 4 graphically compares the outcomes of the three scenarios. The left diagram in Figure 4 (A) shows that as the point allocation becomes steeper, the percentage of multi-member teams increases from 55.0% in the flat-curve scenario, to 61.5% under the current policy, and further to 72.3% in the steep curve scenario. This indicates that the increase of the competitive pressure motivates participants to collaborate. It affects the high-high collaborations the most, as the percentage of this type of teams increases from 8% under the flat-curve scenario to around 20% under the steep-curve scenario (see the right diagram of Figure 4 (A)). High-low collaborations , however, will decline when the competitive pressure is too high. As the result of the increase in collaborations, the top performance will improve. The left diagram of Figure 4 (B) shows that the top performance increases from 3.60 under the flat-curve scenario to 3.69 under the steep-curve scenario. Such increase will benefit the sponsoring business.

All of the above results are similar to those presented in Section 7.1, where we manip-

ulate the informativeness of public tiers. In addition, the right diagram of Figure 4 (B), however, shows that the average performance across all teams will increase from 0.83 under the flat-curve scenario to about 0.97 under the steep-curve scenario. What drives the majority of teams to perform better under the competitive pressure is the large jumps in the number of multi-member teams, as shown in the left diagram of Figure 4 (A). The increase in collaborations makes everyone become more productive. Although the increase in the informativeness of public tiers in the first counterfactual also encourages the formation of teams, by comparison the change is not as drastic as from the increase in the competitive pressure.

Figure 4: Impact of Competitive Pressure



Bars represent 90% confidence intervals estimated based on 100 rounds of simulations.

7.3 Discussion

The mechanisms that affect collaborations and team performances under the two counterfactuals are very different. Improving the informativeness of the public tiers reduces the uncertainty about the ability of teammates and benefits high-ability participants. This comes at the cost of low-ability participants. A steep point-allocation system would create competitive pressure, forcing participants to collaborate to compete effectively. Kaggle could use this insight to design a better policy that creates more value not only for sponsoring businesses but also the participants on the platform.

Our study focuses on Kaggle's competitions, but the findings could be generalized to many other contexts where collaboration is crucial to success. For example, co-authorship among economists has been found to be growing over the years, which helps increase the number

of publications for individual economists (e.g., Hollis, 2001; Ductor, 2015). Collaborations are also prevalent in the service industry. Salespeople work as teams to provide customers better services (e.g., Chan et al., 2014a,b). Financial advisors of banks form teams to combine expertise and provide better account management for clients[19]. In the gaming industry, it is a common strategy for gamers to tackle difficult tasks together in multiplayer games[20]. Collaborations are also frequently observed at the firm level. Intra-firm partnership for R&D projects is crucial for innovation (Un et al., 2010). Collaborative marketing is also a popular strategy that involves working with similar firms to promote brands, minimize costs, and increase sales[21].

We have shown that reducing the uncertainty about potential collaborators will improve the performances of top teams; however, the average performance across teams will decrease. In some business cases (e.g. salesforce and financial advisor teams), the total performance of all teams could be more important than the performance of the best team. Our finding suggests that improving the informativeness of public signals may not be beneficial in these cases. Rather, increasing the competitive pressure through a more concentrated allocation of payoffs to a few top teams can improve the performance of not only the best team but also the majority of teams, as the second counterfactual shows.

The counterfactuals demonstrate that the maximum score increases by 0.04 from the no-information to high-information scenario, and 0.09 from the flat-curve to the steep curve scenario. A potential concern is that the magnitude of the change seems quite small. To better understand the magnitude of the increases, we use the results in Table 9 as a benchmark: when a high-ability individual forms a team with another high-ability individual, the predicted improvement in the performance score is 0.3 relative to when she competes alone, and 0.35 relative to when she forms a team with a medium-ability individual. Compared with these predicted improvements, the changes in the team performance in the two counterfactuals are not negligible.

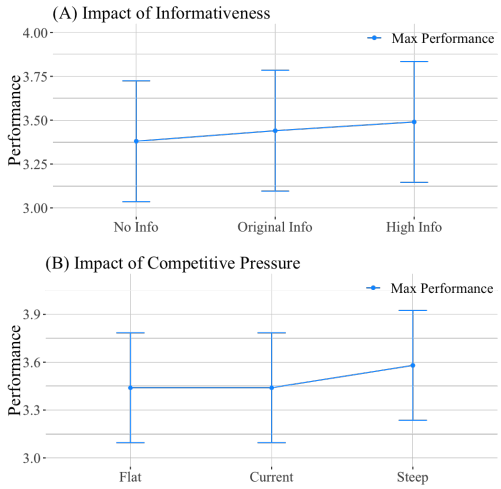
Further improvement in the maximum performance score is difficult to obtain because the score comes from the best team among a large number of high-ability participants (about 640). In cases where talents are not as abundant, the improvement of the top performance will be bigger. To illustrate this point, we create a new competition in which the conditional probabilities of high ability for Novice, Contributor and Expert are reduced. The number of participants and the percentages of four tiers remain the same as in the previous counterfactuals. Compared with the 30%, 23%, and 47% split for the low-, medium-, and high-ability

[19]Source: `https://investmentsandwealth.org/teams`.
[20]Source: `https://www.ediiie.com/blog/gaming-trends-2020`.
[21]Source: `https://www.edge-creative.com/blog/what-are-the-benefits-of-collaborative-marketing`.

participants in the original competition, we increase the proportion of low-ability participants to 49%, decrease the proportion of high-ability participants to 24%, and adjust the medium-ability participants to 27%.

We repeat the two counterfactual exercises. Changes in the maximum performance of the top team, when the informativeness of public tiers and the competitive pressure increase, are shown in Figure 5 (A) and (B). The magnitude of the changes has increased: as public tiers become more informative, the maximum performance increases from 3.38 in the no-information scenario to 3.49 in the high-information scenario. The 0.11 increase is significantly larger than the 0.04 increase in Section 7.1. Similarly, as competitive pressure increases, the maximum performance increases from 3.44 in the flat curve scenario to 3.58 in the steep curve scenario. The 0.14 increase again is larger than the 0.09 increase in Section 7.2.

Figure 5: Performances of Top Teams in New Competition



Bars represent 90% confidence intervals estimated based on 100 rounds of simulations.

One of the assumptions we make in the counterfactuals is that participants do not change across scenarios. As public tiers become more informative and more points are awarded to top teams, the number and composition of participants entering the competition are likely to change, which consequently can influence the equilibrium outcomes. We expect more high-ability and fewer low-ability individuals will enter the competition. As such, the effects on team formation and performance might be stronger than what we showed. Ignoring the entry decisions of participants is a limitation in our counterfactuals.

Since we do not observe how transfers are made between participants in a team, we assume that they split the benefit of forming the team, i.e., α_{ij}, equally independent from their tiers or abilities. We make the assumption because we cannot separately identify α_{ij} and

τ_{ij}. In reality, the split could be asymmetric, so one may be concerned that the assumption can affect our results. To investigate this issue, we repeat the counterfactuals assuming that the split is $\frac{1}{3}$ for the lower-tier individual and $\frac{2}{3}$ for the higher-tier individual. Results show that, as the lower-tier (higher-tier) individual gains less (more) from α, the transfer from the lower-tier individual to the higher-tier individual will reduce by a similar amount. More importantly, changes in the team formation and team performances remain largely the same in the counterfactuals. Therefore, we believe the symmetry assumption of α does not have significant impact on our key findings.

8 Conclusions

Collaboration is a common phenomenon among individuals, within companies, and across industries: gamers tackle difficult tasks together in multiplayer games; salespeople work as teams to provide customers better service; firms work together on designing modularized components in the supply chain. Understanding the motives and outcomes of collaborations and designing marketplaces for efficient collaboration generates significant values.

Using Kaggle as an empirical context, we address two main issues that are neglected in the literature. First, potential participants in collaborations may not fully observe the ability of others. When payoffs are tied with abilities, such uncertainty may impede the incentive of collaboration. We develop a structural matching model that incorporates the incomplete information of participants and show that, when the public signals (i.e., tier status) for abilities are more informative, the incentive for collaboration and the performance of top performers will both increase. Second, individuals collaborate to compete against other collaborations. Our model incorporates the competitive effect in the payoff function where payoff depends on performance rankings. We use counterfactuals to show that incomplete information may hinder collaboration and that high competitive pressure is beneficial to participants, the sponsoring business and the platform because it can boost collaboration and increase collaboration efficiency. The findings illustrate the mechanism of how incomplete information and competition affect collaboration, and they can be generalized to other collaborative markets where incomplete information and competition are important components. We make a methodological contribution to the literature by developing a model that captures a complicated market environment where incomplete information is prevalent, spillover from matching exists, and transfers between collaborators are unobserved. Our model framework in its entirety or simplified forms could be applied to large scale one-to-one matching games that involve numerous participants.

We made a few simplifying assumptions to keep the model tractable. Future research

could relax these assumptions to further understand the underlying mechanism that drives collaborations. First, we model collaborations as one-to-one matching. When the collaboration involves more participants, the problem will become more complicated. Recent research that studies network formation (e.g., Ho and Lee, 2017; Ghili, 2022) offers an alternative way to model such type of collaborations. Second, our model treats entry of participants as exogenous. To control for the potential issue that participants with different abilities may choose competitions based on monetary reward, we allow the distribution of ability types to be conditional on monetary reward. Future research can study how the entry decision may affect collaborations, if more granular data such as click streams are readily available. Third, we assume that transfers will not affect the team performance. This is a necessary assumption for the model identification because we do not observe the transfer from the data. We combine various important economic factors, such as the economy of scale, complementarity of skills, and moral hazard in a reduced-form way in the team performance function that is specified on vertical differentiation of abilities. Future research could further investigate how these factors separately affect the incentive and outcomes of collaborations. Fourth, we assume perfect information on the performance function for all tiers of participants. Allowing the information on performance function to differ among participants will require us to incorporate learning in the model, which is infeasible in our setting. We also acknowledge the limitation that we can not exactly tell if team members had other interaction that provides better-quality signals. Future research could further investigate the implications from learning on collaboration. Finally, one important assumption for our matching model is the large market assumption. Adapting our model to matching markets with small number of participants may require extra efforts to capture preference heterogeneity in a continuous way.

References

Ackerberg, Daniel A, Maristella Botticini. 2002. Endogenous matching and the empirical determinants of contract form. *Journal of Political Economy* **110**(3) 564–591.

Agrawal, Ajay, Christian Catalini, Avi Goldfarb. 2015. Crowdfunding: Geography, social networks, and the timing of investment decisions. *Journal of Economics & Management Strategy* **24**(2) 253–274.

Baccara, Mariagiovanna, Ayşe İmrohoroğlu, Alistair J Wilson, Leeat Yariv. 2012. A field study on matching with network externalities. *American Economic Review* **102**(5) 1773–1804.

Baik, Kyung Hwan. 2016. Endogenous group formation in contests: Unobservable sharing rules. *Journal of Economics & Management Strategy* **25**(2) 400–419.

Bajari, Patrick, Jeremy T Fox. 2005. Measuring the efficiency of an fcc spectrum auction. Tech. rep., National Bureau of Economic Research.

Bamford, James, David Ernst, David G Fubini, et al. 2004. Launching a world-class joint venture. *Harvard Business Review* **82**(2) 90–100.

Banerjee, Suryapratim, Hideo Konishi, Tayfun Sönmez. 2001. Core in a simple coalition formation game. *Social Choice and Welfare* **18**(1) 135–153.

Bayus, Barry L. 2013. Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management Science* **59**(1) 226–244.

Becker, Gary S. 1973. A theory of marriage: Part I. *Journal of Political Economy* **81**(4) 813–846.

Burtch, Gordon, Anindya Ghose, Sunil Wattal. 2013. An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets. *Information Systems Research* **24**(3) 499–519.

Chan, Tat Y, Barton H Hamilton, Nicholas W Papageorge. 2015. Health, risky behaviour and the value of medical innovation for infectious disease. *The Review of Economic Studies* **83**(4) 1465–1510.

Chan, Tat Y, Jia Li, Lamar Pierce. 2014a. Compensation and peer effects in competing sales teams. *Management Science* **60**(8) 1965–1984.

Chan, Tat Y, Jia Li, Lamar Pierce. 2014b. Learning from peers: Knowledge transfer and sales force productivity growth. *Marketing Science* **33**(4) 463–484.

Choo, Eugene. 2015. Dynamic marriage matching: An empirical framework. *Econometrica* **83**(4) 1373–1423.

Choo, Eugene, Aloysius Siow. 2006. Who marries whom and why. *Journal of Political Economy* **114**(1) 175–201.

Dagsvik, John K. 2000. Aggregation in matching markets. *International Economic Review* **41**(1) 27–58.

DasGupta, Anirban. 2008. *Asymptotic theory of statistics and probability*. Springer Science & Business Media.

Druedahl, Louise C, Timo Minssen, W Nicholson Price. 2021. Collaboration in times of crisis: a study on COVID-19 vaccine R&D partnerships. *Vaccine* **39**(42) 6291–6295.

Ductor, Lorenzo. 2015. Does co-authorship lead to higher academic productivity? *Oxford Bulletin of Economics and Statistics* **77**(3) 385–407.

Echenique, Federico. 2008. What matchings can be stable? the testable implications of matching theory. *Mathematics of Operations Research* **33**(3) 757–768.

Echenique, Federico, M Bumin Yenmez. 2007. A solution to matching with preferences over colleagues. *Games and Economic Behavior* **59**(1) 46–71.

Eriksson, Tor. 1999. Executive compensation and tournament theory: Empirical tests on Danish data. *Journal of Labor Economics* **17**(2) 262–280.

Farrell, Joseph, Suzanne Scotchmer. 1988. Partnerships. *The Quarterly Journal of Economics* **103**(2) 279–297.

Fox, Jeremy T. 2008. Estimating matching games with transfers. Tech. rep., National Bureau of Economic Research.

Fox, Jeremy T. 2010. Identification in matching games. *Quantitative Economics* **1**(2) 203–254.

Fox, Jeremy T, Patrick Bajari. 2013. Measuring the efficiency of an FCC spectrum auction. *American Economic Journal: Microeconomics* **5**(1) 100–146.

Fox, Jeremy T, David H Hsu, Chenyu Yang. 2012. Unobserved heterogeneity in matching games with an application to venture capital. Tech. rep., National Bureau of Economic Research.

Gale, David, Lloyd S Shapley. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly* **69**(1) 9–15.

Galichon, Alfred, Bernard Salanié. 2021. Cupid's invisible hand: Social surplus and identification in matching models. *arXiv preprint arXiv:2106.02371* .

Ghili, Soheil. 2022. Network formation and bargaining in vertical markets: The case of narrow networks in health insurance. *Marketing Science* **41**(3) 501–527.

Graham, Bryan. 2011. Econometric methods for the analysis of assignment problems in the presence of complementarity and social spillovers. *Handbook of Social Economics* **1** 965–1052.

Ho, Kate, Robin S Lee. 2017. Insurer competition in health care markets. *Econometrica* **85**(2) 379–417.

Hollis, Aidan. 2001. Co-authorship and the output of academic economists. *Labour Economics* **8**(4) 503–530.

Holmstrom, Bengt. 1982. Moral Hazard in Teams. *The Bell Journal of Economics* **13**(2) 324.

Huang, Yan, Param Vir Singh, Kannan Srinivasan. 2014. Crowdsourcing new product ideas under consumer learning. *Management Science* **60**(9) 2138–2159.

Kini, Omesh, Ryan Williams. 2012. Tournament incentives, firm risk, and corporate policies. *Journal of Financial Economics* **103**(2) 350–376.

Koopmans, Tjalling C, Martin Beckmann. 1957. Assignment problems and the location of economic activities. *Econometrica: Journal of the Econometric Society* 53–76.

Kräkel, Matthias, Gunter Steiner. 2001. Equal sharing in partnerships? *Economics Letters* **73**(1) 105–109.

Lazear, Edward P. 1989. Pay equality and industrial politics. *Journal of Political Economy* **97**(3) 561–580.

Lazear, Edward P, Sherwin Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* **89**(5) 841–864.

Levin, Jonathan, Steven Tadelis. 2005. Profit sharing and the role of professional partnerships. *The Quarterly Journal of Economics* **120**(1) 131–171.

Liu, Qingmin, George J Mailath, Andrew Postlewaite, Larry Samuelson. 2014. Stable matching with incomplete information. *Econometrica* **82**(2) 541–587.

Mindruta, Denisa. 2013. Value creation in university-firm research collaborations: A matching approach. *Strategic Management Journal* **34**(6) 644–665.

Ni, Jian, Kannan Srinivasan. 2015. Matching in the sourcing market: A structural analysis of the upstream channel. *Marketing Science* **34**(5) 722–738.

Pycia, Marek. 2012. Stability and preference alignment in matching and coalition formation. *Econometrica* **80**(1) 323–362.

Roth, Alvin E, Marilda Sotomayor. 1992. Two-sided matching. *Handbook of Game Theory with Economic Applications* **1** 485–541.

Sørensen, Morten. 2007. How smart is smart money? a two-sided matching model of venture capital. *The Journal of Finance* **62**(6) 2725–2762.

Stephen, Andrew T, Peter Pal Zubcsek, Jacob Goldenberg. 2016. Lower connectivity is better: The effects of network structure on redundancy of ideas and customer innovativeness in interdependent ideation tasks. *Journal of Marketing Research* **53**(2) 263–279.

Uetake, Kosuke, Yasutora Watanabe. 2017. Entry by merger: Estimates from a two-sided matching model with externalities. *Available at SSRN 2188581* .

Un, C Annique, Alvaro Cuervo-Cazurra, Kazuhiro Asakawa. 2010. R&d collaborations and product innovation. *Journal of Product Innovation Management* **27**(5) 673–689.

von Siemens, Ferdinand A, Michael Kosfeld. 2014. Team production in competitive labor markets with adverse selection. *European Economic Review* **68** 181–198.

Wilson, Alistair, Mariagiovanna Baccara, Ayse Imrohoroglu, Leeat Yariv. 2010. A field study on matching with network externalities. *Proceedings of the Behavioral and Quantitative Game Theory: Conference on Future Directions*. ACM, 96.

Wu, Chunhua. 2015. Matching value and market design in online advertising networks: An empirical analysis. *Marketing Science* **34**(6) 906–921.

Yang, Yupin, Mengze Shi, Avi Goldfarb. 2009. Estimating the value of brand alliances in professional team sports. *Marketing Science* **28**(6) 1095–1111.

Yi, Sang-Seung. 1997. Stable coalition structures with externalities. *Games and Economic Behavior* **20**(2) 201–237.

Yoganarasimhan, Hema. 2015. Estimation of beauty contest auctions. *Marketing Science* **35**(1) 27–54.

Online Appendix

A Calculation of Team Performance Rank

We now derive the probability of team performance rank for a team $Pr(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}|A_i, A_j, \mathcal{M}))$.The rank of a team is determined by its own performance and the performances of the other teams. Since the performance of a team is driven by the true ability of team members, we first need to calculate team structure $\mathcal{M}$ in terms of teams' true ability. We define a team type t by the true ability of team members, $t = \langle a, a'\rangle$ for multi-member teams and $t = \langle a, \emptyset\rangle$ for single-member teams. Given A types of true ability for participant, we have a total of $T = \frac{A(A+1)}{2} + 2A$ unique team types.

Given $\mathcal{M}$, we could calculate the percentage of team type $Pr_T(t)$ with the following equation.

$$Pr_T(t) = \sum_{r\in R}\sum_{r'\in R} Pr_R(r)\cdot Pr_R(r')\cdot Pr(a|r,r')\cdot Pr(a'|r',r). \tag{A.1}$$

where $t = \langle a, a'\rangle$, $Pr_R(r)$ and $Pr_R(r')$ represent the proportion of participants with signal r and r', and $Pr(a|r,r')$ represents the updated probability of a participant's true ability a conditional on her own signal r and her choice of teammate of signal r', as defined in equation (10) in Section 4.3. The team structure $\mathcal{M}$ could then be characterized by the proportions of team type $Pr_T(t)$ for all $t \in T$.

In Section 4.2, we assume the performance of a team with type $t = \langle a, a'\rangle$, such that $Y(t)$ follows normal distribution $N(\lambda_t, \sigma_\xi^2)$. The performance of one team $Y \in Y_{\mathcal{M}}$ follows a mixture normal distribution, with each of the underlying component to be distributed as $N(\lambda_t, \sigma_\xi^2)$and the probability of each components to be $Pr_T(t)$. Based on the property of mixed normal distribution, the cumulative distribution function of team performance Y under $\mathcal{M}$ is defined as

$$F_Y(y) = \sum_{t\in T} Pr_T(t)\Phi(y, \lambda_t, \sigma_\xi^2). \tag{A.2}$$

and the probability density function of Y is defined as

$$f_Y(y) = \sum_{t\in T} Pr_T(t)\phi(y, \lambda_t, \sigma_\xi^2) \tag{A.3}$$

We use $Y_{(p)}$ to represent the p_{th} order statistics of Y and $Pr_T(p, t|Y_{(p)} = y)$ to represent the probability that the p_{th} order statistic is from a particular team type t conditional on the p_{th} order statistics of Y equals y, then the value of $Pr_T(p, t|Y_{(p)} = y)$ could be derived

using Bayesian Rule

$$Pr_T(p,t|Y_{(p)}=y)=\frac{\phi(y|\lambda_t,\sigma_\xi^2)Pr_T(t)}{\sum_{t'\in T}\phi(y|\lambda_{t'},\sigma_\xi^2)Pr_T(t')} \tag{A.4}$$

Then we integrate $Pr_T(p,t|Y_{(p)}=y)$ over the distribution of order statistics $Y_{(p)}$ and get the unconditional probability that the p_{th} order statistic is from a particular team type t, $Pr_T(p,t)$ as

$$Pr_T(p,t)=\int Pr_T(p,t|Y_{(p)}=y)f_{Y_{(p)}}(y)dy \tag{A.5}$$

Finally, the probability that a specific team $\langle i,j\rangle$ with type $t=\langle A_i,A_j\rangle$ ranks the p_{th}, $Pr(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}|A_i,A_j)=p)$ is equal to the probability that p_{th} order statistic is from team type t divided by the number of teams with the same team type t, which is the total number of teams M times the proportion of team type t, i.e.,

$$Pr(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}|A_i,A_j,\mathcal{M})=p)=\frac{Pr_T(p,t=\langle A_i,A_j\rangle)}{M\times Pr_T(t=\langle A_i,A_j\rangle)}. \tag{A.6}$$

The challenge of calculating this probability comes from the complicated form of the exact distribution of the order statistics $Y_{(p)}$ for mixture normal distribution. We utilize the property of the asymptotic distribution of the order statistic function for mixture normal distribution to help alleviate the computational burden (DasGupta, 2008). Specifically,

$$Y_{(p)}\sim N\left(F_Y^{-1}(\frac{p}{M}),\frac{\frac{p}{M}(1-\frac{p}{M})}{M[f_Y(F_Y^{-1}(\frac{p}{M}))]^2}\right). \tag{A.7}$$

where $\frac{p}{M}$ is the specific quantile that defines the p_{th} order, and F and f are cumulative distribution and density function of Y defined in equations (A.2) and (A.3). We can simulate values from this asymptotic distribution and compute the numerical integration of equation (A.5).

Specifically, the procedure of the expected probability calculation is outlined as follows:

1. Simulate S_1 random numbers ν_1 from the standard normal distribution, and simulate S_2 random numbers ν_2 from the standard normal distribution.

2. Given the model parameters λ,σ_ξ, compute the proportion of team types $Pr_T(t)$ in equation (A.1), then separately scale $Pr_T(t)\times S_1$ samples of ν_1 to be $\nu_t=\lambda_t+\sigma_\xi\nu_1$ for each team type t. This gives us the mixture normal distribution of team ability according to the team structure.

3. Rank-order the above values ν_t, and numerically compute the quantile function F_Y^{-1}.

4. For each rank p, compute the following:

   (a) Compute the mean and variance of the asymptotic normal distribution specified in equation (A.7), and scale the S_2- generated standard normal random numbers ν_2 to ν_p with the calculated mean and variance of $Y_{(p)}$;

   (b) Use the simulated random numbers ν_p for each team type t to compute the numerical integration in equation (A.5); and

   (c) Compute the probability for team performance rank in equation (A.6) for each team type t.

B Proof: Existence of Equilibrium

As explained in the paper, the equilibrium $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ is characterized by the fixed point of the system of equations $\boldsymbol{\mathcal{H}} : (\boldsymbol{Pr}(\boldsymbol{\mathcal{M}}), \boldsymbol{\tau}) \to (\boldsymbol{Pr}(\boldsymbol{\mathcal{M}}), \boldsymbol{\tau})$. So the existence of equilibrium is equivalent to the existence of a fixed point for $\boldsymbol{\mathcal{H}}$. The proof is done in two steps. First, we show that in equilibrium transfer is finite, so we could restrict the domain of $\boldsymbol{\mathcal{H}}$, $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ to be a compact and convex subset of the Euclidean space. Second, we show that $\boldsymbol{\mathcal{H}}$ mapped from $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ onto itself is continuous. Therefore, we can use Brouwer's fixed point theorem on $\boldsymbol{\mathcal{H}}$ to prove the existence of a fixed point.

Proof. Because $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})$ is the matching probability for a participant of R signals and A true abilities with participants of R signals, the coordinates of $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})$ forms a vector in a vector space of $R \times R \times A$ dimension. $\boldsymbol{\tau}$ is the transfer between participants with different signals. The coordinates of $\boldsymbol{\tau}$ is a vector in $\frac{R\times(R+1)}{2}$ vector space. Because we assume both R and A are finite, the coordinates of $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ is a vector in $(R \times R \times A + \frac{R\times(R+1)}{2})$ dimension vector space. $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ is a point in Euclidean space of dimension $(R \times R \times A + \frac{R\times(R+1)}{2})$.

Suppose the set of $\boldsymbol{\tau}$ is unbounded, $\exists r, r'$, s.t. $\tau(r, r') = +\infty$, $\tau(r', r) = -\infty$. $\forall a, a'$, $Pr(\langle i, j\rangle | a, r, r'; \tau) = 1$, $Pr(\langle i, j\rangle | a', r', r; \tau) = 0$. $\forall Pr(A|R)$, $Pr(\langle r, r'\rangle|\tau) = 1$, $Pr(\langle r', r\rangle|\tau) = 0$, market equilibrium constraint is not satisfied. Thus in equilibrium the set of $\boldsymbol{\tau}$ is bounded and there exists a finite number B, s.t. each coordinate of $\boldsymbol{\tau}$ is in the finite interval $[-B, B]$. The set of $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})$ is bounded and closed because each coordinate of $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})$ is a probability that lies in the unit interval of $[0, 1]$. We restrict $\mathcal{D}$, the domain of $\boldsymbol{\mathcal{H}}$ to be a closed and bounded subset of Euclidean space. Because each coordinate of $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ is in a closed and bounded interval, the convex combination of two points in $\mathcal{D}$ is still in $\mathcal{D}$, i.e $\mathcal{D}$ is convex. Based on the specification in the paper, each member function of $\boldsymbol{\mathcal{H}}$ is continuous, and thus $\boldsymbol{\mathcal{H}}$ is continuous. $\forall(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}}), \boldsymbol{\tau}) \in \mathcal{D}$, $\boldsymbol{\mathcal{H}}(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}}), \boldsymbol{\tau}) \in \mathcal{D}$, because $\boldsymbol{h_1}$ yields a mapping from a set of probabilities on to itself and $\boldsymbol{h_2}$ comes from the market equilibrium

constraint that controls the boundary of $\boldsymbol{\tau}$. Thus $\boldsymbol{\mathcal{H}}$ is a continuous function from a compact and convex set $\mathcal{D}$ onto itself.

By Brouwer fixed point theorem, the fixed point $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ exists.□

C Simulation of Symmetric Benefits in Teams

In the model, we make the symmetric benefit assumption, i.e. $\alpha(R_i, R_j) = \alpha(R_j, R_i)$ for each unique combination of (R_i, R_j). Now we use simulations to show that the assumption is necessary for identification.

We construct two sets of parameters. One is the original parameters in the Result section estimated under the symmetric benefit assumption, and the other is the set of modified parameters with everything remaining the same except using asymmetric benefits. The two sets of benefits are shown in Table C.1.

Table C.1: Benefits from Forming Teams

Team Structure	Novice	Contributor	Expert	Master
Symmetric Benefits				
Novice	151	162	183	220
Contributor	162	276	289	310
Expert	183	289	378	407
Master	220	310	407	552
Asymmetric Benefits				
Novice	151	262	183	120
Contributor	62	276	289	310
Expert	183	289	378	407
Master	320	310	407	552

Specifically, in the asymmetric benefit scenario, we increase $\alpha_{Novice,Contributor}$ by 100 while decreasing $\alpha_{Contributor,Novice}$ by 100, so the sum $(\alpha_{Novice,Contributor} + \alpha_{Contributor,Novice})$ remains the same as that in the symmetric benefit scenario. We also decrease $\alpha_{Novice,Master}$ by 100 while increasing $\alpha_{Master,Novice}$ by 100.

Then using each set of parameters, we simulate the equilibrium outcomes and recover the transfers by solving for the fixed point in equation (15) for each competition in the data.

Table C.2 shows the average transfer τ from one to another tier across competitions between different tiers in the two scenarios. Compared with the τ in the symmetric-benefit scenario, on average, $\tau_{Novice,Contributor}$ decreases by 111, and $\tau_{Novice,Master}$ increases by 109. Using the recovered $\tau_{Novice,Contributor}$'s in 101 competitions, we run a T-test and find that the decrease in $\tau_{Novice,Contributor}$ is not significantly different from 100 (p value = 0.65). Similarly,

the increase in $\tau_{Novice,Master}$ is not significantly different from 100 (p value = 0.10). For the rest of the α's, the differences between the two scenarios are not significantly different from 0.

Table C.2: Average Transfer Between Participants

Paid by	To Teammate			
	Novice	Contributor	Expert	Master
Symmetric Benefits				
Novice	0	2701	3477	4231
Contributor		0	1088	3046
Expert			0	1449
Master				0
Asymmetric Benefits				
Novice	0	2590	3475	4393
Contributor		0	1093	3069
Expert			0	1477
Master				0

Note: Transfer for participants with the same tier is 0. Positive values in the upper triangle mean that participants of lower tier will pay positive transfer to participants in higher tier.

We further compare the equilibrium outcomes in terms of the average percentage and average performance of each team type. The differences between the two scenarios are very close to 0. The results are reported in Table C.3 and Table C.4.

Table C.3: Difference in Team Types between Two Scenarios

Tier/Choice	Single	Team			
		Novice	Contributor	Expert	Master
Novice	-0.000021	-0.0001	-0.00012	0.000065	0.000086
Contributor	-0.00028		-0.000052	0.000044	-0.000037
Expert	-0.00017			0.000039	0.00019
Master	-0.00036				0.00031

Note: Each value is calculated as the estimated result in the asymmetric benefit scenario minus that in the symmetric benefit scenario.

The simulation results suggest that the changes in the α after relaxing the symmetric assumption could be absorbed by the corresponding changes in τ such that the sum $\alpha_{R_i,R_j} + \tau_{R_i,R_j}$ remains the same for each pair of (R_i, R_j). Since the equilibrium outcomes in the two scenarios are also not significantly different, we conclude that only the sum $\alpha_{R_i,R_j} + \tau_{R_i,R_j}$ could be identified from the data. Thus, the symmetric benefit assumption is necessary for identification.

Table C.4: Difference in Team Scores between Two Scenarios

Tier/Choice	Single	Team			
		Novice	Contributor	Expert	Master
Novice	0.00015	0.00085	-0.0028	-0.0013	0.0029
Contributor	-0.0014		-0.0025	0.014	-0.015
Expert	-0.0012			0.0059	-0.0012
Master	0.0033				0.0068

Note: Each value is calculated as the estimated result in the asymmetric benefit scenario minus that in the symmetric benefit scenario.

D Simulation of Unique Equilibrium

Proposition 1 proves the existence of equilibrium. In this sub-section, we use simulations to numerically show that the equilibrium is unique in the ranges of parameters that are resonable for our empirical context.

We first generate 50 sets of parameters Θ by randomly drawing from a sequences of uniform distributions. The ranges are $[-5, 5]$ for the team performances λ, $[0, 20]$ for the preferences θ, $[-2000, 2000]$ for the benefits in teams α, $[0, 2000]$ for the utility scale μ, $[0, 1]$ for the standard deviation of performance δ_ξ, and $[0, 1]$ for the condition probability $Pr(A|R)$. We believe these ranges are reasonable, and wide enough for the estimation.

For each set of parameters, we randomly select one competition and numerically solve for the equilibrium $(\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})^*, \boldsymbol{\tau}^*)$ using 20 sets of different starting values. We randomly select one equilibrium outcome as benchmark, and compute the difference between the benchmark and each of the other 19 equilibrium outcomes.

Finally, we combine all the differences between $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})$ into a large vector of length $50 \times 19 \times 60$, where 60 is the number of elements in $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})$.

Figure D.1 shows the density plot of the vector of differences in $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}})$. We find a large mass of differences centering around 0. The absolute value of more than 90% of all the 57,000 differences is less than 0.01.[22] This indicates that the vast majority of equilibrium outcomes $\boldsymbol{Pr}(\boldsymbol{\mathcal{M}}^*)$ numerically converge to the same value. We find similar results for the equilibrium $\boldsymbol{\tau}^*$. Thus, we numerically show that in the reasonable ranges of parameters, the equilibrium is unique.

[22]The convergence criterion in the inner loop solving for the fixed point is set as 0.01; the differences that are close to 1 are usually caused by the equilibrium where team formation probability is 0 for all the team types, but such equilibrium does not appear in the data.

Figure D.1: Density of the Differences between Equilibrium Outcomes