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

Tat Y. Chan*
Olin Business School, Washington University in St. Louis, chan@wustl.edu

Yijun Chen
Olin Business School, Washington University in St. Louis, yijun.chen@wustl.edu

Chunhua Wu
Sauder School of Business,University of British Columbia chunhua.wu@sauder.ubc.ca

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 behind collaborations, with 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 of participants that determine whether and with whom they form teams. Using model parameters, we conduct policy experiments to investigate how the collaboration efficiency is affected by the incomplete information and competitive pressure on the platform. Our results provide implications on how firms could better align individual incentives to foster and improve collaborations.

Key words: Matching game, Collaboration, Incomplete information

1. Introduction

Collaboration is ubiquitous. It plays a critical role in enhancing productivity among academic researchers. For example, co-authorship among economists has been found to be growing over years, which helps increase the number of publications for individual economists (e.g., Hollis 2001, Ductor 2015). Collaboration also helps accelerate business innovations and new product development in various industries. Bamford et al. (2004) documented that more than 5,000 joint ventures, and many more contractual alliances, have been launched worldwide in the past five years of 2004. Yet they also found that only about half of the joint ventures could achieve returns greater than the investment cost. They argued that having incompatible partners is an important reason for the failures. Collaboration is also an important determinant of employee performance within firms, as

* The authors thank seminar participants at Purdue University, Washington University, Frontiers Empirical Marketing Conference and the 2017 Marketing Science Conference for their valuable feedbacks.

numerous industry studies have shown that workplace collaboration is a key factor for a company's success.[1]

There are multiple benefits from collaboration, including the facilitation of the economy of scale, complementarity of knowledge and skills, and division of labor, that help tackle complicated work tasks. Other factors may facilitate or impede collaborations. In particular, potential participants in collaboration may have heterogeneous abilities and skills that are not fully observed by other participants. Such lack of information can lead to adverse selection and other inefficiencies, as identified in the economics literature. Another important factor is competition. Because the payoff for being the first can be much higher than being in other ranks, academic researchers race to find a breakthrough for scientific problems, firms seek to launch new products earlier than their competitors, and employees compete to be the best among peer workers. Under such competitive pressure, the collaboration of other parties may force an individual also to collaborate. When there are a large number of participants with heterogeneous abilities and skills competing in the market, whether and with whom to collaborate becomes a very intricate problem.

We develop a structural one-sided matching model in this paper to study how individuals collaborate to compete. The model has several unique features. First, participants compete against each other in the market, and the success of their collaboration effort can reduce the returns of other participants in the market. Second, there is a large number of participants with heterogeneous abilities in the matching game. Their abilities are partially reflected by the imperfect, public information. Finally, our model allows potential collaborators to negotiate how rewards and costs are shared. Collaboration will only be successful if all of the involving parties agree on the sharing rule. To analyze the properties of this model, we focus on the market equilibrium, which is characterized by each participant's optimal choice regarding whether to collaborate and with whom to collaborate, under two constraints. The first constraint is that each individual makes rational inferences on the true ability of other participants based on the public information and their collaboration decisions, using a Bayesian updating framework. The second constraint imposes that the sharing of rewards and costs agreed by collaborators will clear the market. That is, the number of one type of participants, defined by the public information, who want to match with another type of participants, is equal to the number of the latter type who want to match with the former. This equilibrium concept is first developed in the theory paper of Becker (1973) who studied the marriage market. We extended his model to allow for incomplete information on the ability of other parties, and competition among collaborations. We prove the existence of such an equilibrium in a large-scale, one-sided matching game.

[1] Source: https://www.smartsheet.com/how-workplace-collaboration-can-change-your-company.

We use this matching model to study how collaborations affect the individual performance and competition outcomes. We also investigate what is the role of incomplete information and competition in the formation of collaborations and, based on the results, what policies a firm can use to enhance the efficiency of collaborations. Given the prevalence of collaborations among firms and individuals, the answers to those questions are of high economic importance. We apply our model to a dataset we collect from Kaggle.com, a leading data science competition platform. There are several reasons why the empirical context is suitable for our study. First, Kaggle connects firms that provide data and sponsors competitions with participants (i.e.,data scientists) who provide solutions in order to win competitions. Monetary and non-monetary rewards from competitions are based on the ranking of the team performance, a format equivalent to rank-order tournaments (Lazear and Rosen (1981)). Second, each competition typically attracts hundreds or even thousands of participants. To improve their performance, participants may form teams (i.e.,collaborations) to compete against the others. Third, there is a fixed policy on how Kaggle points are allocated to each participant in collaborations. How to split the monetary reward and how to share the workload, on the other hand, are negotiated by the participants when forming teams. Finally, as repeated interactions across competitions are rare (only 9% of all team interactions in data), a participant is unlikely to have full information regarding the true ability or skills of potential collaborators. Kaggle makes the information on the "tier" status and Kaggle points accumulated from past competitions of each participant publicly available on its website; however, such information is not perfectly aligned with the true ability. This last feature makes it challenging to incentivize participants to collaborate. Kaggle makes the information on the "tier" status and Kaggle points accumulated from past competitions of each participant publicly available on its website; however, these information are not perfectly align with the true ability. This last feature makes it challenging to incentivize participants to collaborate. We find from data only about one-fourth of participates collaborated in competitions, even after Kaggle changed its policy of awarding Kaggle points to encourage collaborations.

Estimating the structural matching model is challenging because of two issues: first, due to the competition nature the likelihood for an individual to form teams is a function of the likelihoods that other participants will form teams. Second, how team members split the monetary reward and workload is unobserved to researchers. Because of these two reasons, the likelihood function cannot be evaluated analytically. To tackle this problem, we propose a two-level estimation procedure. Conditional on a given set of parameters at the outer level, we impose the equilibrium constraints in the inner level. This methodology can be applied to estimate other types of large-scale matching games when competition or other forms of spillovers exist, sharing-rule is not predetermined, or imperfect information exists.

Estimation results show that the tier status of a participant, a piece of publicly available information, can reasonably reflect her true ability. However, there is a large variation in ability across participants who belong to the same tier, suggesting that the information is a noisy signal. 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 teammates with lower ability are negative, implying the risk of collaborations that is due to the lack of complete information. Finally, we recover the market clearing sharing rule between participants when the market is at equilibrium. We find that participants at a lower tier have to pay a positive (monetary and non-monetary) transfer to teammates at a higher tier. This explains why a significant proportion of high tier participants are willing to form teams with participants from lower tiers in data.

After recovering the model primitives, we conduct counterfactual analyses to investigate how the incomplete information and competition affect collaborations and their outcomes, as well as what policies Kaggle may use to enhance the efficiency of collaborations. In the first counterfactual, we manipulate the informativeness of the tier status regarding the true ability of individuals. We find that, in comparison to the case when the tier status is uninformative, improving the informativeness will increase the maximum performance from all teams, a performance measure that the sponsoring business for the data science competition care most, and the expected payoff of participants. The increases in participants' performance and expected payoffs are critical for Kaggle's business because, as a platform, its success heavily relies on the ability of attracting sponsoring businesses and top talents on both sides.

The second counterfactual studies how the extent of competition affects collaboration outcomes. Specifically, Kaggle can manipulate the degree of competitiveness by changing how its non-monetary Kaggle points are allocated to teams at different ranks. Fixing the aggregate number of points in each competition, the competitive pressure increases when more points are awarded to high-ranked teams at the expense of teams at lower ranks. When points are equally allocated, there will be no competition (except for the monetary reward). We manipulate the point allocation function in the counterfactual, and find that increasing the competitive pressure for Kaggle points will boost collaboration among participants, as well as improve the best performance of all teams. It also has positive effect on the payoffs of top participants; however, the effect on the payoff of general participants is negative.

Combining results from the two counterfactuals, we conclude that, to attract talents and improve team performances, Kaggle should focus on providing more information about the true ability of participants. Whether Kaggle should make the point allocation more competitive depends on the objectives of the platform. If the platform wants to improve the best performance, a high competitive pressure for points is preferred. However, if attracting more participation in competitions is the priority, Kaggle should avoid awarding points to the top few teams only.

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. These two factors have not been taken into account in the traditional matching literature. Our matching model also allows for unobserved sharing rules between agents, a factor that is not accounted for by the previous literature of coalition games. We also develop a method for estimating our matching game. The method could be easily applied to other empirical settings where collaborations are critical for the success in market competition. It can be adopted to study other markets where incomplete information exists and matching has spillover effects. For substantive contributions, this paper provides insights on how collaborations affect the individual performance and competition outcomes, and how information and competition affect the collaboration efficiency. Our counterfactuals generate implications on what policies firms may use to enhance collaborations.

The rest of the paper is organized as follows. We discuss the related literature in Section 2, then describe the empirical context and provide some summary statistics in Section 3. Section 4 develops the matching model. Detailed model specification, identification and the estimation are presented in Section 5. Estimation results and counterfactual analyses are presented in Section 6. Finally, we discuss the model 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 have explored 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 recent studies in different industries (e.g., Fox and Bajari 2013, Yang et al. 2009, Wu 2015). A few recent papers study the vertical matching between insurance networks and hospitals using the matching model. Ho and Lee (2017), for example, uses a Nash-in-Nash framework as the equilibrium concept in the matching game. A similar modeling approach is also adopted in Ghili (2018). 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 is first developed in Becker (1973), and is later adopted in the empirical work of Choo and Siow (2006) that studies 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. They also make the market clearing assumption so that they can estimate the model using the maximum likelihood estimator with equilibrium constraints. Our model differentiates 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 Pycia 2012, Farrell and Scotchmer 1988). 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 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 for a digital journal and test several economic theories using substitution models and reinforcement models. In another study, Bayus (2013) studies individual ideators' 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 for generating high potential ideas, but they are relatively slow for learning the cost of implementation. The above research on crowdsourcing has treated collaborations on platforms as exogeneously given. We contribute to the

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

literature by studying how participants of Kaggle's competitions collaborate and how their outcomes are affected by collaboration.

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 rewards based on ranking could motivate hard work and improve performance. For example, Eriksson (1999) uses the compensations for executives to test the tournament theory. Kini and Williams (2012) finds 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 previous works by investigating how collaborations can enhance the team performance, and thus how the competitive environment in rank-order tournaments can increase the incentive to collaborate.

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 talents. On the demand side, firms provide data for the business problems they want to solve or opportunities they want to explore. On the supply side, data scientists, researchers and students who have the talents 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 set up by sponsoring businesses. By the end of 2017, Kaggle has hosted 248 competitions, attracted more than 60 thousand participants, and awarded over 9 million US dollars. These competitions have resulted in significant scientific advancements including furthering the state of art in HIV research, improving predictive technologies and algorithms, and uplifting operational efficiency in business applications.[3]

For most competitions, the sponsoring business first specifies the winning rules and monetary prizes. Depending on the business background, type of analytics required and the amount of the prize, each competition attracts a distinct set of participants. They can compete alone or form

[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.

teams. It is very common that thousands of participants 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, and the awarded points decline as a convex function with lower ranks. This point allocation creates an additional competitive pressure among participants, on the top of the competition for the monetary prize. Kaggle also uses a tier system to classify individuals, under which participants who have the highest Kaggle points accumulated from past competitions are awarded the Master status, followed by Expert and then Contributor tiers. Participants who compete for the first time are recognized as Novices.[4] The tier status and points can be an important part of the non-monetary reward for participants. As Kaggle has gradually established its reputation in the data science community, showing the tier status or points is a useful way to strengthen the resume of data scientists. During an interview with Wired Magazine, Gilberto Titericz, a top Kaggle player, claimed that job opportunities that flow from a good Kaggle ranking are generally more bankable than money prizes.[5]

Collaborations can be important for participants to achieve good performance and win competitions. To make sure that winners from competitions can provide high-quality solutions to sponsoring businesses, Kaggle designs rules for participants that not only allow, but also encourage participants to collaborate. Indeed, the formal unit of participation is a "team", where a participant competing alone is just a "single-member team." Collaborations on Kaggle are formed in a decentralized way, in which participants decide whether to form teams and with whom to form team by themselves, usually through an invitation-and-acceptance/rejection procedure. Mutual agreements among team members are needed but, once the team is formed, it is not allowed to change throughout the competition.

Despite of the potential benefits of collaboration, there are many factors that may deter team formation. The first factor is the lack of information about other participants. Kaggle tries to solve this problem by making the tier status of each participant publicly available on the website. The information nevertheless is an imperfect measure. For example, the abilities for new participants (who are all Novices) are not well distinguished, and participants who have participated in more competitions are more likely to belong to a high tier. Furthermore, the sharing rule for the monetary and non-monetary rewards may discourage high ability participants to collaborate with others whom

[4] For a more detailed description on the points allocations and tier progressions, see an article at: https://www.kaggle.com/progression. Some of the 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/. Accessed January 20, 2018.

they are not familiar with. The monetary prize if a team wins has to be split between members in a way that is 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 each participant in a team may not be higher than that from competing alone.

To encourage collaboration, Kaggle changed the point allocation policy in 2016. Before the change, the points a member could get is equal to what her team wins divided by the team size. The new policy divides the points of the team by the square root of the team size.[6] Simultaneously, Kaggle also reduced the number of points a team can win at each rank. Single-member teams therefore can win 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[7] for the empirical application. The dataset includes information on competitions, participants, teams and the final score, ranking, and rewards for each team. We observe 315 competitions that cover a time span of 7 years from 2010 to 2016. We exclude competitions that do not award Kaggle points, which are designed to let participants "have fun" and familiarize with the competition. We also exclude competitions that have less than 100 participants. These are mostly competitions at the very early stage of Kaggle when it launched in 2010. Since the value of Kaggle points were not well recognized by the data scientist community, the incentives for participants could be very different from the competitions in later stage when Kaggle points are highly valued. After excluding these two types of competitions, we are left with 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, which we will show later.

Table 3.2 provides the summary statistics on several key variables at the competition level. The total monetary prizes in our data is US $25,000 averaged across competitions, with the highest at US $500,000. A competition attracts 643 participants on average. In general, competitions with higher monetary rewards attract more participants.

[6] Source: http://blog.kaggle.com/2015/05/13/improved-kaggle-rankings.

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

Table 1 Monetary Rewards And Participants Across Competitions

Rewards Quartile	Rewards (USD)			Participants		
	min	mean	max	min	mean	max
Q1	0	360	950	27	226	779
Q2	1000	4675	9000	26	463	1883
Q3	10000	11283	17500	32	627	4151
Q4	20000	67621	500000	97	1374	6260

Table 2 shows how participants of the 4 Kaggle tiers differ in their Kaggle points. We report the average points per competition that a participant joined before, and the total accumulative points. Novices have not participated in any competition and thus have 0 points. As a participant's tier moves up, both dimensions of Kaggle points also increase. The last column of the table shows that more than 50% of participants are Novices. Masters are an elite group, as only 10% of participants belong to this tier.

Table 2 Summary Statistics of Participants' Tier and Kaggle Points

Player Tier	Mean Average Points per Competition	Mean Total Points	No. of Participants
Novice	0.00	0.00	34929
Contributor	845	1425	10924
Expert	1702	5876	6679
Master	4371	22179	5721

Note: Participants may join multiple competitions, so the total number of participants in this table is larger than 31,246(unique participants across competitions).

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, probably because of different reasons. Novices form teams in order to learn and prove her ability by collaborating with others. Masters, on the other hand, are well recognized in the community for their high abilities, and thus they are highly demanded for collaborations. Second, there is a pattern of sorting, as 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.

We now look at how collaborations impact the performance. In almost all of the competitions, performance is measured by the predictive accuracy on hold-out samples, but the criteria used for calculating the accuracy differs from competition to competition.[8] Since the measure is unique for

[8] Some of the most commonly used evaluation algorithms are Root Mean Squared Errors (RMSE), Root Mean Squared Errors (RMSE), Root Mean Squared Logarithmic Error (RMSLE), Area Under Receiver Operating Characteristic Curve (AUC), and Log Loss.

Table 3 Summary Statistics of Participants' 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%	2.9%	9.7%	4.0%
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.

each competition, we create a standardized score from the original performance measure in each competition. We first calculate the mean and standard deviation of the original performance measure for single-member teams. We deduct the original performance of each team by the calculated mean and then 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 could be compared across competitions.

The average and the standard deviation of the standardized score across different types of teams are reported in Table 4. Several interesting patterns arise. First, conditional on a participant's tier, her performance when forming team is in general better than when she competes alone. Since the performance provides value for sponsoring businesses, this result suggests that Kaggle should encourage more collaborations, a policy it seems to have long adopted. A natural question one may ask is, if this is the case, why there are still a large percentage of participants who stay single, as shown in first column of Table 3? One of the major reasons could be that, since participants have to split the monetary prize and Kaggle points and face the potential conflicts due to moral hazard and free-riding, the expected payoff of each member when forming team could be lower than if she competes alone. Another data pattern is that teaming with a participant with a higher tier status will perform better than teaming with another with a lower status. Specifically, the average performance is the highest when teaming with a Master. Finally, there is a large variation in performances within each type of single- or multiple-member teams, suggesting that there is a large differentiation in the ability of participants even belong to the same tier. This is especially interesting for Novices. The average performance of those who team up with Experts and Masters is very high, but the performance of those who team up with other Novices or Contributors are relatively lower. This implies that the ability of Novices is very heterogeneous. The heterogeneity in participants' ability brings uncertainty in expected payoff in teams and these may deter participants' team formation.

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.

4. Model

In this section, we develop a structural matching model, explicitly incorporating the incomplete information and competition, to study the outcomes of the matching when the market is at equilibrium. We model an agent's team formation decision as a one-sided matching game with a large number of individuals, and apply this model to the Kaggle competition. Our modeling approach can be easily generalized to a broader setting where collaborations are important in order to compete with other agents.

Below, we will formalize the model in four steps. First we describe the information set and the payoff function of a participant when she forms team with another participant. Next, we discuss how we model the expected monetary and non-monetary rewards. We then explain how participants form expectations conditional on the information set. Finally, we introduce the equilibrium concept and explain how the equilibrium can be represented by the participant's optimal choice of whether and with whom to form team under the rational expectations and market clearing constraints.

4.1. Model Setup and the Payoff Function

For each competition, define the set of participants in the competition as $\mathcal{N}$, and the number of participants as N. Also define 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.

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 made a few additional assumptions. First, we assume participants when forming teams will negotiate how the monetary reward and team work should be divided. The agreement cannot be broken once the team is formed. Potential issues from forming teams, including moral hazard and resulted 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, but in the model estimation we approximate how the pool of participants may change in different competitions. Third, we treat each competition as a static game, so that we can focus on the determinants of team formation within games and ignore the strategic dynamic interactions between participants across games. Finally, we also treat the monetary and non-monetary rewards pre-specified in the competition as exogenous.

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 a 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 for other participants to identify the focal participant's 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 a very informative signal. The distribution of signals is informative for the abilities of the participant population if, for any two participants i and j, $Pr\left(A_i = a|R_i\right) \neq Pr\left(A_j = a|R_j\right)$ for $R_i \neq R_j$. The conditional probability is also assumed to be common knowledge.

The performance of team $\langle i,j\rangle$ is determined by the abilities A_i and A_j, denoted as $Y_{\langle i,j\rangle}(A_i, A_j)$. We assume that, first, team performance independent from the ordering of i and j, i.e., $Y_{\langle i,j\rangle}(A_i, A_j) = Y_{\langle j,i\rangle}(A_j, A_i)$. Second, team performance does not depend on the abilities of participants in other teams. However, the rank of $Y_{\langle i,j\rangle}$ will depend on the performance of all 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 rank 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 solve its business problem. Participants, on the other hand, care about the ranking since it determines how much the monetary reward, denoted by $Money(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}$, and how much the non-monetary reward (i.e., Kaggle points), denoted by $Point(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}$, the team can earn from the competition (details are in Section 4.2).

Kaggle decides how the Kaggle points awarded to a team should 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 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 reason, 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 positive 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 of the monetary reward will belong to the participant, i.e., $\gamma^M(R_i,\varnothing,\mathcal{R})=1$, where "$\varnothing$" indicates that the target participant does not exist in team $\langle i,\varnothing\rangle$. For the share of team work, we use $\tau(R_i,R_j,\mathcal{R})$ to denote the "transfer" of workload from i to j (relative to equal share of the work). As a participant's agreement to handle more of the workload implies 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 is sum up to zero. The transfer of a single-member team is normalized to zero, i.e.,$\tau(R_i,\varnothing,\mathcal{R})=0$.

Finally, there are additional benefits from team works, including the economy of scale and specialization in job tasks, 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 the top of how team works can impact the performance in the competition; such impact has been captured in $Y_{\langle i,j\rangle}(A_i,A_j)$. We cannot separately identify these additional benefits and costs from data; therefore, our model only incorporates the net benefit from the above factors. To allow for the heterogeneity of the net benefit across teams, we assume that it is determined by the types of the team members defined by the public signals. That is, the net benefit is represented by a function $\alpha(R_i,R_j)$. To simplify the analysis we assume the function is independent from the ordering of the abilities, i.e., $\alpha(R_i,R_j)=\alpha(R_j,R_i)$. We also normalize the net benefit of competing solo to zero, i.e., $\alpha(R_i,\varnothing)=0$.

Combining the above components, we assume that, when the focal participant i is considering the collaboration with the target participant j, she will form expectation of her payoff relative to all of the other team formation 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 function, 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 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 $R_j=R_{j'}$.[9]

[9] 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).

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 makes decision on which type of individuals she should team up with. Assuming that ε_{i,R_j} is distributed as Type-I extreme value distribution with scale parameter μ. Given a sharing rule γ^M and a transfer rule τ, the probability that the participant's optimal choice is teaming with a participant with signal r (including single-member team with $r = \varnothing$), can be calculated as:

$$\Pr_{\mathcal{M}}\left(A_i, R_i, r, \mathcal{R} | \gamma^M, \tau\right) = \frac{\exp\left(V\left(A_i, R_i, r, \mathcal{R}; \gamma^M, \tau\right)/\mu\right)}{\sum_{r' \in (R \cup \varnothing)} \exp\left(V\left(A_i, R_i, r', \mathcal{R}; \gamma^M, \tau\right)/\mu\right)} \tag{3}$$

where the subscript "$\mathcal{M}$" on the left side denotes the matching probability, and $V\left(A_i, R_i, r, \mathcal{R}; \gamma^M, \tau\right)$ on the right side is the expected payoff in equation (1) (or equation (2) if $r = \varnothing$) without the random component ε_{i,R_j}. Furthermore, R_j in the equation is replaced by r, and $\gamma^M\left(R_i, R_j, \mathcal{R}\right)$ and $\tau\left(R_i, R_j, \mathcal{R}\right)$ by γ^M and τ, respectively.

Note that, first of all, γ^M and τ are unobserved by researchers. To evaluate the choice probability we will have to back out these variables from equilibrium conditions. Second, the expectations of the monetary and non-monetary rewards in equations (1) and (2) are over the true abilities of the target participant as well as that of all other participants in the competition. The focal participant will make inference on the abilities of other participants based on the public signals,as well as their decisions on how to form teams with other types of participants, when the market is at equilibrium. These two issues will be discussed in Section 4.3.

4.2. Performance and Rewards

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

$$Y_{\langle i,j\rangle}\left(a, a'\right) = \lambda'_{aa} + \xi_{ij} \tag{4}$$

where $\lambda_{aa'}$ is a model parameter to be estimated representing the predicted performance of a team with ability a and a'. By definition $\lambda_{aa'} = \lambda_{a'a}$.[10] The stochastic term ξ_{ij} captures other unobserved factors that affect the final performance, and is assumed to distribute 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.

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

The collection of performances of all teams under team structure $\mathcal{M}$ is $Y_{\mathcal{M}}$. The expected 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 $Pr\left(A_j = a|R_j, R_i, \mathcal{R}; \gamma^M, \tau\right)$ be the probability that the true ability of target participant j, whose signal is R_j, conditional on the focal participant's signal is R_i, the collection of signals of all participants in the competition denoted by $\mathcal{R}$, and sharing rule γ^M and transfer τ. This conditional probability also represents the updated belief of participant i over j's ability, 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 updated belief in the next sub-section.

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

$$\begin{aligned}&E\left[Money\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right)\right)|A_i, R_i, R_j, \mathcal{R}; \gamma^M, \tau\right] = \sum_{p=1}^{P}\left[Prize_p \times \Pr\left(Z_{(ij)}\left(Y_{\mathcal{M}}\right) = p\right)\right] \\ &= \sum_{p=1}^{P}\left[Prize_p \times \sum_{a\in A}\Pr\left(A_j = a|R_j, R_i, \mathcal{R}; \gamma^M, \tau\right) \times \Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}|A_i, a, \mathcal{M}\right) = p\right)\right]\end{aligned} \tag{5}$$

In the above equation, the probability on the right side in the first line denotes the probability that the rank of team $\langle i,j\rangle$ is at the p^{th} place. 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, the expected non-monetary reward (i.e., Kaggle points) for team $\langle i,j\rangle$ can be specified as

$$\begin{aligned}&E\left[Point\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right)\right)|A_i, R_i, R_j, \mathcal{R}; \gamma^M, \tau\right] = \sum_{p=1}^{M}\left[Point_p \times Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}\right) = p\right)\right] \\ &= \sum_{p=1}^{M}\left[Point_p \times \sum_{a\in A}Pr\left(A_j = a|R_j, R_i, \mathcal{R}; \gamma^M, \tau\right) \times Pr\left(Z_{\langle i,j\rangle}\left(Y_{\mathcal{M}}|A_i, a, \mathcal{M}\right) = p\right)\right]\end{aligned} \tag{6}$$

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 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 (5) and (6). The computation is complicated because it involves a rank order distribution. We use the asymptotic 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. Using 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.[11]

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 the following:

$$Pr\left(A_j=a|R_j,R_i,\mathcal{R};1-\gamma^M,-\tau\right)=\frac{Pr_{\mathcal{M}}\left(a,R_j,R_i,\mathcal{R}|1-\gamma^M,-\tau\right)\times Pr\left(A_j=a|R_j\right)}{\sum_{a'\in A}Pr_{\mathcal{M}}\left(a',R_j,R_i,\mathcal{R}|1-\gamma^M,-\tau\right)\times Pr\left(A_j=a'|R_j\right)} \quad (7)$$

where $\Pr_{\mathcal{M}}\left(a,R_j,R_i,\mathcal{R}|1-\gamma^M,-\tau\right)$ is j's choice probability given that her true ability is a, as defined in equation (3). 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 (7) implies rational expectation in the updated belief because the belief is based on j's optimal choice.

Until now the probability of i choosing a teammate with signal R_j and the resulted expected monetary and non-monetary rewards are all conditional on a specific sharing rule γ^M and transfer τ (and $1-\gamma^M$ and $-\tau$ for j). Researchers do not observe γ^M and τ. 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 round. γ^M and τ have to satisfy this market-clearing condition. As γ^M and τ cannot be separately identified, 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, 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 $(\widetilde{\gamma}^M-1/2)$ multiplied by the expected monetary reward. The choice probability will remain unchanged.

With the normalization, let $Pr_M(r,r'|\mathcal{R},\tau)$ 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

$$Pr_{\mathcal{M}}\left(r,r'|\mathcal{R},\tau\right)=\sum_{a\in A}Pr_{\mathcal{M}}\left(a,r,r',\mathcal{R}|\gamma^M=1/2,\tau\right)\times Pr(A=a|r) \quad (8)$$

where $Pr_M(a,r,r',R|\gamma^M=12,\tau)$ is defined in equation (3). 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:

$$Pr_{\mathcal{M}}\left(r,r'|\mathcal{R},\tau\left(r,r'\right)\right)\times Pr_R(r)=Pr_{\mathcal{M}}\left(r',r|\mathcal{R},-\tau\left(r,r'\right)\right)\times Pr_R\left(r'\right) \quad (9)$$

[11] 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.

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

Substitute equation (8) into (9), and further plug equation (3) into the equation, then apply logarithmic transformation and move terms, we can derive that

$$\begin{aligned}\tau(r,r') =& \frac{\mu}{2} \cdot [\ln Pr_R(r') + \ln \left(\sum_a \frac{exp(V(a,r',r,\mathcal{R};\gamma^M = 1/2, \tau(r,r')) + \tau(r,r')/\mu)}{\sum_{\tilde{r}\in(R\cup\varnothing)} exp(V(a,r',\tilde{r},\mathcal{R};\gamma^M = 1/2, -\tau(r,\tilde{r}))/\mu)} \times Pr(A = a|r') \right) \\ & - \ln Pr_R(r) - \ln \left(\sum_a \frac{exp(V(a,r,r',\mathcal{R};\gamma^M = 1/2, \tau(r,r')) - \tau(r,r')/\mu)}{\sum_{\tilde{r}\in(R\cup\varnothing)} exp(V(a,r,\tilde{r},\mathcal{R};\gamma^M = 1/2, \tau(r,\tilde{r}))/\mu)} \times Pr(A = a|r) \right)]\end{aligned} \tag{10}$$

This expression helps us to prove the existence of the market equilibrium, which is in the next sub-section.

4.4. Market Equilibrium

The matching game in our model is characterized by the preference parameters $\{\theta_i^M, \theta_i^P, \alpha(R_i, R_j)\}$ for every participant (see equations (1) and (2)), the monetary and non-monetary rewards $\{Prize_p, Point_p\}$ for every rank (see equations (5) and (6)), 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 (9) 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 probability $Pr_{\mathcal{M}}$ and transfer τ.

Let $\boldsymbol{Pr^*_{\mathcal{M}}}$ be a $AR(R+1) \times 1$ vector that represents the collection of the choice probabilities of all ability and signal types (including single-member team choice), and $\boldsymbol{\tau^*}$ be a $(R^2+1) \times 1$ vector that represents the collection of transfers from one to another signal type within a team (the transfer in single-member team is fixed to zero). We can combine equations (3) and (10) into a system of equations $\boldsymbol{\mathcal{H}: (Pr_{\mathcal{M}}, \tau) \rightarrow (Pr_{\mathcal{M}}, \tau)}$:

$$\begin{cases} \boldsymbol{Pr_{\mathcal{M}} = h_1(Pr_{\mathcal{M}}, \tau)}, \\ \boldsymbol{\tau = h_2(Pr_{\mathcal{M}}, \tau)}, \end{cases} \tag{11}$$

Market equilibrium $\boldsymbol{(Pr^*_{\mathcal{M}}, \tau^*)}$ is the solutions of the equation system $\boldsymbol{\mathcal{H}}$.

Proposition 1. *For each competition characterized by* $\{\theta_i^M, \theta_i^P, \alpha(R_i, R_j)\}$ *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 (11) exists.*

The proof is in Appendix B.

5. Model Estimation

When estimating the model using data from Kaggle, we discretize ability into three types, i.e., $A = \{Low, Medium, High\}$. We use the tier status to proxy the noisy signal. That is, $R = \{Novice, Contributor, Expert, Master\}$. Although other signals, such as a participant's accumulated Kaggle points, are available, they are highly correlated with the tier status. The tier is also the most highlighted part when checking a participant's profile. Therefore, it should be the most important signal for a participant's ability.

For model parameters, we normalize the marginal utility of the monetary reward θ_i^M to 1, and allow the marginal utility of the non-monetary reward θ_i^P to differentiate based on the participant's tier status. That is, $\theta^P = \{\theta^P_{Novice}, \theta^P_{Contributor}, \theta^P_{Expert}, \theta^P_{Master}\}$. Such heterogeneity captures the fact that a participant's need for Kaggle points may differ at different tiers. We also allow θ^P to change before and after Kaggle adjusted its point allocation system. It reflects the fact that we find from data the points a participant can earn from a competition are significantly different after the policy change.

We allow $\alpha(R_i, R_j)$ in equation (1) to differ across each unique combination of (R_i, R_j), but assume that $\alpha(R_i, R_j) = \alpha(R_j, R_i)$. Consequently, there are 10 α's to be estimated (as α's for single-member teams are normalized to zero). Since we do not observe the actual ability, the probability $Pr(A_i = a|R_i)$ in equation (7) is estimated from data. This probability is specific for every unique combination of ability and tier; therefore, the number of probabilities is $4 \times 3 - 4 = 8$. As a reduced-form way of capturing how competitions with various prize levels may attract different pools of talents to participate, we also allow the probabilities to be different for competitions with small and large monetary rewards.

For the team output function in equation (4), we estimate the λ for each unique combination of (a_i, a_j), as well as the λ for single-member teams of each ability level. Therefore, there are 9 λ's to be estimated. Finally, we also estimate the variance δ_ξ^2 and the scale parameter μ in equation (3).

5.1. Maximum Likelihood under Equilibrium Constraints

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

$$\begin{aligned}
&\mathcal{L}(\langle i,j\rangle | R_i, R_j, \mathcal{R}, \tau) = Pr_{\mathcal{M}}(R_i, R_j|\mathcal{R}, \tau) \times Pr_{\mathcal{M}}(R_j, R_i|\mathcal{R}, -\tau) \\
=&(\sum_{a\in A} Pr_{\mathcal{M}}(a, R_i, R_j, \mathcal{R}|\gamma^M = 1/2, \tau) \times Pr(A_i = a|R_i))\times \\
&(\sum_{a'\in A} Pr_{\mathcal{M}}(a', R_j, R_i, \mathcal{R}|\gamma^M = 1/2, -\tau) \times Pr(A_j = a'|R_j))
\end{aligned} \tag{12}$$

The equation indicates that the team will only be formed if it is optimal for both participants.[12]

Let $\phi(y, \lambda(a,a'), \delta_\xi^2)$ be the (normal) pdf of the performance of the team when the abilities of the participants members are a and a', defined in equation (4). The likelihood that the team performance is $Y_{\langle i,j\rangle}$ is

$$\begin{aligned}&\mathcal{L}(Y_{\langle i,j\rangle}|R_i, R_j, \mathcal{R}, \tau) = \sum_{a\in A}\sum_{a'\in A} Pr_{\mathcal{M}}(a, R_i, R_j, \mathcal{R}|\gamma^M = 1/2, \tau)\cdot \\ &Pr_{\mathcal{M}}(a', R_j, R_i, \mathcal{R}|\gamma^M = 1/2, -\tau)\cdot \phi(Y_{\langle i,j\rangle}, \lambda(a,a'), \delta_\xi^2)\end{aligned} \tag{13}$$

The likelihood function we use in model estimation is the sum of the log likelihood of the observed teams and their performance in every competition in data. That is,

$$l(\Theta) = \sum_g \sum_{\langle i,j\rangle\in\mathcal{M}_g} [l(\langle i,j\rangle|R_i, R_j, \mathcal{R}, \tau) + l(Y_{\langle i,j\rangle}|R_i, R_j, \mathcal{R}, \tau)] \tag{14}$$

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 $Y_{\langle i,j\rangle}|R_i, R_j, \mathcal{R}, \tau)$ are the log functions of the likelihoods in equations (12) and (13), respectively.

The challenges of evaluating equation (14) are two-fold. First, $\boldsymbol{\tau}$ is unobserved to researchers; it has to be recovered from the market-clearing condition. Second, the matching probability $\boldsymbol{Pr_{\mathcal{M}}}$ in equations (12) and (13) comes from equation (3). Because of competition, the payoff function of forming teams depends on how other teams in the competition are formed. This means that the matching probability is a function of the matching probabilities of other teams in the competition, as the first line offunction $\boldsymbol{h_1}$ in equation (11) suggests. Because of these two issues, the likelihood function cannot be evaluated analytically.

We propose a two-level estimation procedure to tackle this problem. In the inner level, conditional on trial parameters Θ we search for the matching probabilities and transfers $\boldsymbol{(Pr^*_{\mathcal{M}}, \tau^*)}$ for every competition such that $\boldsymbol{Pr^*_{\mathcal{M}} = h_1(Pr^*_{\mathcal{M}}, \tau^*)}$ and $\boldsymbol{\tau^* = h_2(Pr^*_{\mathcal{M}}, \tau^*)}$. That is, we find $\boldsymbol{(Pr^*_{\mathcal{M}}, \tau^*)}$ that satisfy the equilibrium constraints. In the outer level, we search for Θ that maximizes the likelihood function in equation (14). The detailed algorithm is the following:

1. Start with initial value Θ^0. For any trial parameters Θ,

(a) Assume initial value $\boldsymbol{\mathcal{H}(Pr^0_{\mathcal{M}}, \tau^0)}$. Calculate the expected payoffs, using numerical methods;

(b) Calculate $\boldsymbol{(Pr'_{\mathcal{M}}, \tau') = \mathcal{H}(Pr^0_{\mathcal{M}}, \tau^0)}$

(c) Replace $\boldsymbol{\mathcal{H}(Pr^0_{\mathcal{M}}, \tau^0)}$ by $\boldsymbol{(Pr'_{\mathcal{M}}, \tau')}$. Repeat the above procedure until $\boldsymbol{(Pr_{\mathcal{M}}, \tau)}$ converge. They represent the market equilibrium under model parameters Θ^0

[12] For a single team $\langle i, \varnothing\rangle$, the second component on the right side is fixed to 1.

2. Calculate the likelihood function value in equation (14) under parameters Θ. Search for Θ such that the likelihood function is maximized.

Note that Proposition 1 proves the existence but not the uniqueness of $(\boldsymbol{Pr}^*_{\boldsymbol{\mathcal{M}}}, \boldsymbol{\tau}^*)$. The potential multiple equilibria are a concern when we estimate the model and conduct counterfactuals. During the estimation, we test whether this is an issue by varying the starting value $(\boldsymbol{Pr}^0_{\boldsymbol{\mathcal{M}}}, \boldsymbol{\tau}^0)$ 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.

5.2. Identification

The identification of the unobserved team ability distribution (λ, δ^2_ξ), and the public tier to private ability conditional distribution $Pr(A|R)$, comes from how team score changes under different combinations of R_i and R_j, as equation (13) suggests. In the likelihood function, $\mathcal{L}(Y_{\langle i,j \rangle})$ can be treated as a latent class regression, with $Pr(A|R)$ represents the size of latent classes.

The proportions of teams collaborated by different types of 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 $\langle R_i, R_j \rangle$ indicates the larger the value of $\alpha\langle R_i, R_j \rangle$, relative to $\alpha\langle R_i, \varnothing \rangle$ that is 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. If, for example, collaborations increase the points expected to win more than competing solo, a large number of collaborations across competitions will indicate a high θ^P. After the policy change, the share of Kaggle points a participant can obtain from forming teams increases, relative to competing solo. Therefore, participants with higher θ^P will be more likely to form teams. This policy change also helps the identification of the parameter.

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. Since the marginal utility of the monetary prize is normalized to one, the lack of variation will imply a high value of μ.

At the inner level of the model identification, we also treat the equilibrium $(\boldsymbol{Pr}^*_{\boldsymbol{\mathcal{M}}}, \boldsymbol{\tau}^*)$ as model parameters. The identification of these parameters comes from the equilibrium constraints, represented by equations (3) and (10).

6. Result

In this section, we first report the estimation results. Based on the results, we then use counterfactuals to explore how Kaggle can increase the value for both sponsoring businesses and individual participants on both sides, through providing better information about participants' ability and manipulating the competitive pressure through the point allocation policy.

6.1. Estimation Results

Estimated marginal utilities for Kaggle points (θ^P) are reported in Table 5. As the marginal utility for the monetary prize is normalized to one, the estimates in the table represent how much a participant is willing to pay for 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 estimate the marginal utilities before and after the policy change separately.

Table 5 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 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 competitions. In data, more than 80% of Novices only participated in one competition. Their marginal utility for points thus may be low. The rest 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 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. Another interesting result is that marginal utilities for Kaggle points increase after the policy change. One of the possible reasons 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 – 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 very valuable for participants. Use the estimates multiplied by the average number of points a participant wins in competitions, 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 the non-monetary payoff dominates the monetary payoff.

Table 6 reports the estimated net benefits, on the top of gaining more Kaggle points and the monetary reward, from forming teams (i.e., $\alpha(R, R')$) All of the estimates are significantly positive, implying that the benefits of forming teams dominate 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 collaboration are important. Furthermore, the net benefits from teaming with individuals with high tiers are higher than that with low tiers. For example, the benefits for an Expert from working with a Master are $407, about twice as high as working with a Novice ($183). Such a high value may come from the benefit of learning from high ability peers.

Table 6 Parameter Estimates of Preferences in 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 7 reports the estimated conditional probability of belonging to an ability level given a participant's tier status (i.e., $Pr(A|R)$). Since competitions that offer higher prizes may attract more talented individuals, we group the competitions in our data into low- and high-prize type, 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 and a high-prize one attracts about 1,200 participants. We estimate the conditional probabilities for these two types of competitions separately. Results show that Kaggle's tier system is in general consistent with participants' true ability. For example, the proportion of individuals with high ability increases from 33-38% for Novices to 90-94% for Masters. However, the variation in abilities within each tier is very substantial, implying that tiers are a noisy signal. This is especially true for Novices, as the proportions of individuals with low and high ability are both large. Finally, we calculate the unconditional probabilities of each ability type by summing up the products of the conditional probability and the size of each tier, across all tiers. They are reported in the rows of "Total". In comparison with small-prize competitions, the proportion of individuals with high ability is larger in high-prize competitions while that with low and medium ability is smaller, suggesting that high monetary rewards are able to attract more talents.

Table 7 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	0.28	0.26	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 8 reports the estimated productivity, which is also the average performance (i.e., $\lambda(a, a'))$, of each team combination. Estimates in the last column of the table are the productivity of single-member teams, which can be used as the benchmark against the performance from collaborations. There is a strong increasing productivity from low to high ability single-member teams. Comparing the left columns with the last column, collaborations clearly help improve the team productivity. 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 danger for high-ability participants: if they work alone, the predicted performance is 1.82, higher than that if they team up with medium- or low-ability individuals. This difference is probably due to the division of job tasks, as such poor 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, reducing the noise in tier signals may help improve the effectiveness of collaborations, a result we will show in the counterfactuals.

Finally, the estimated variance of the team productivity (δ_{xi}^2) is 0.53, and the scale parameter in participants' utility function (μ) is 1,313. Both are quite large when comparing with the predicted team performance and average utility level respectively.

From estimation results, we could recover the transfers τ within each competition. The average transfer τ from one to another tier across competitions between different tiers is reported in Table 9. Lower tiers have to pay a positive transfer to higher tiers when forming teams, and the magnitude

Table 8 Parameter Estimates for Team Abilities

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.

increases as the difference in tiers increases. For instance, to form team with a Master, an Expert on average needs to pay \$1,449, while a Novice needs to pay \$4,231. This is because a high-ability teammate will greatly improve the team performance, and other teammates will be benefited from the increased monetary reward and more importantly the non-monetary rewards. Note that the transfers are much higher than the average expected monetary rewards in competitions (\$39), as the non-monetary payoffs are much bigger than the monetary rewards.

Table 9 Average Transfer 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 with the same tier is 0. Negative values in the upper triangle mean that participants of lower tier will pay positive transfer to participants in higher tier.

Our model also allows us to recover the choice probability of participants given the transfer in equilibrium. We find that participants of different abilities have different choice probabilities even if they belong to the same tier. For instance, the probability of low ability Novices teaming with another Novice is 42%, with a Master is 1%, and staying single is 43%. The probabilities for a high-ability Novice are 6%, 4% and 88%, respectively. A high-ability Novice is more likely to stay single, because the loss from splitting points with the other teammate is large. However, if she can team up with another high-ability teammate, her team can have a high chance to achieve a top rank and thus win both monetary and non-monetary rewards. Therefore, she is more likely to collaborate with a Master.

Finally, Table 10 reports the model fit in terms of the average percentage of team types observed in data and that predicted by our model. Overall, the predicted team formations are highly consistent with the data pattern. The only collaboration that the model significantly under-predicts is Masters teaming with Masters. Table 11 compares the average score across each type of teams in data and that predicted in our model. The numbers are again quite close with each other. The model is able to replicate how, for example, by collaborating with a teammate of the same or different tier, a Master is able to obtain an average score higher than competing alone. Overall, our model is able to predict very well team formations and performances as observed in the data.

Table 10 Model Fit: Collaboration Probabilities

Tier/Choice	Single	Team			
		Novice	Contributor	Expert	Master
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%)

Note: Numbers represent percentage of team type in data; Numbers in parenthesis represent predicted percentage of team type from estimation.

Table 11 Model Fit: Team Scores

Tier/Choice	Single	Team			
		Novice	Contributor	Expert	Master
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: Numbers represent average performance for team types in data; Numbers in parenthesis represent predicted average performance from estimation.

7. Counterfactuals

As a platform, the profit of Kaggle relies on the participation of sponsoring businesses on one side and individual talents on the other side. Kaggle has to provide sufficient value to attract both parties. For individual participants, the value of joining competitions is captured by the utility function. For sponsoring businesses, they look for the best solution provided by participants for their business problems. The maximum performance of all teams in a competition would be a good proxy for such value. In this sub-section, we conduct counterfactuals using the estimation results to explore how Kaggle can improve the value it offers to both parties. We focus on two type of policies that Kaggle can implement. First, Kaggle can change the informativeness of the tier system. Second, Kaggle can manipulate the competitive pressure by adjusting the point allocation policy. As we have discussed above, the lack of information can negatively impact the willingness of high-ability individuals to form teams, and that Kaggle points are of great value to participants, the counterfactual policies we explore should have substantial impacts on Kaggle's business.

We use four measures to quantify the impacts: the average and maximum utility of all participants, and the average and maximum performance of all teams, in a competition. The first two measures center around the welfare of participants, while the latter two are directly linked to the value for sponsoring businesses.

7.1. Informativeness of the Tier System

The first counterfactual studies the impacts of changing the informativeness of the tier system. In current practice, Kaggle assigns tiers based on the cumulative Kaggle points a participant has won in past competitions. This is a noisy signal, as Table 2 shows, since individuals who have participated in more competitions are more likely to be assigned a high tier. Furthermore, past performances may not indicate how good an individual will be if the current competition requires a different skill set. Considering one counterfactual scenario that Kaggle only uses the number of past competitions an individual has participated to determine the tier. The tier will be less informative about an individual's true ability than the current practice. In another scenario, suppose Kaggle uses a more sophisticated method to predict the individual's performance for each competition. For example, Kaggle may assign more weights for past competitions that have similar tasks as the current competition. It can also incorporate an individual's current job, college majors, and other relevant attributes as predictors. Doing so the tier status of the individual may vary from one competition to another, and it will be a more informative signal about the individual's ability in a competition.

In the counterfactual, we assume the distributions of participants in terms of the true ability are 1/2, 1/3, 1/6 for Low, Medium and High types, respectively. We also fix the distributions of tiers as 1/3, 1/4, 13/60 and 1/5, for Novices, Contributors, Experts and Masters, respectively. We assume the total number of participants is 1500 and the total money reward is $10,000, and Kaggle's point allocation policy is before the change. We explore three counterfactual scenarios: no information (e.g., the practice that assigns tier based on the number of competitions an individual has participated), low information (e.g., assigning tier based on the cumulative Kaggle points an individual has earned) and high information (e.g., assigning tier based on the sophisticated method discussed above). The information structure of each scenario is specified in Table 12.

For the no-information scenario, the probability of belonging to an ability type is the same across tiers. For the low-information scenario, the probability of belonging to the low-ability type is much higher for Novices and Contributors, while the probability of belonging to the high-ability type is much higher for Masters. Still, there is a high chance that individuals in each tier belong to the other ability types. For the last high-information scenario, Novices and Contributors predominantly belong to low ability, Experts to medium ability, and Masters to high ability. Therefore, the tier status is a more informative signal.

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

Public Type	No Information			Low Information			High Information		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
Novice	0.50	0.33	0.17	0.67	0.25	0.08	0.83	0.13	0.04
Contributor	0.50	0.33	0.17	0.58	0.34	0.08	0.67	0.29	0.04
Expert	0.50	0.33	0.17	0.45	0.54	0.01	0.18	0.78	0.04
Master	0.50	0.33	0.17	0.17	0.25	0.58	0.08	0.25	0.67

Table 13 reports the results from the three scenarios. The first two rows show the percentage of multi-member teams and the percentage of teams that are formed by high ability members. The results suggest that providing better information on the true ability helps facilitate collaborations and, more importantly, increase the chance that high ability participants form teams among themselves. For the participant welfare, the average utility of participants of all types increase from $1,341 under the no-information scenario to$1,511 under the high-information scenario, a 12.7% improvement. The maximum utility among all participants (who comes from a high ability individual) has improved even more by 21.4%. For team performance, the maximum performance among all teams (who comes from a team with two high ability participants) has increased from 3.38 under the no-information scenario to 3.46 under the high-information scenario, a 2.4% improvement. Although this increase seems small, it comes from the best performance team among a large number of participants. Improvement in such a measure is more difficult to obtain. This performance improvement can bring a high value to the sponsoring business. For example, a small increase in the accuracy of demand prediction can help a firm cut down inventory costs and avoid stock-outs and thus significantly improve its profit. Finally, the average team performance has a much more significant increase by 20.4%.

Table 13 Market Outcomes Under Different Information Structure

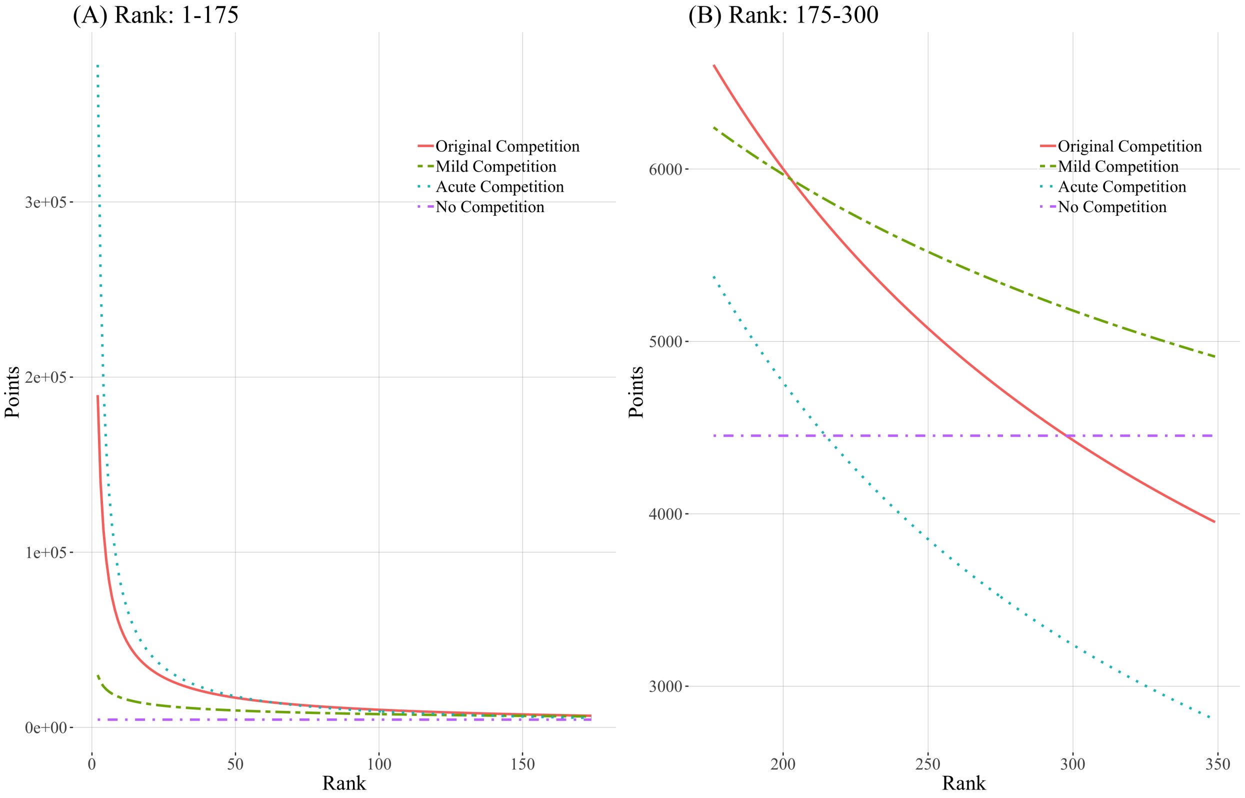
Scenario	No Information	Low Information	High Information
Percentage of Multi-Member Teams	59%	65%	66%
Percentage of High-High Collaborations	6%	10%	12%
Mean Utility	1341	1492	1511
Max Utility	31350	35921	38087
Mean Performance	0.49	0.56	0.59
Max Performance	3.38	3.44	3.46

Overall, the results suggest that improving the informativeness of the tier status can bring significant value for both sponsoring businesses and participants. This creates a win-win situation for both sides, which is the key for the success of Kaggle's business.

7.2. Point Allocation

Our second counterfactual studies how the competition for the non-monetary reward, i.e., Kaggle points, affects the participant welfare and team performances. We vary the extent of competition by adjusting the slope of the point allocation policy. A flat slope means points are allocated more evenly across teams, while a steep slope puts more weight on the performance ranking. We use the average points difference between two neighboring performance ranks as the measure of the slope, which is equivalent to the average absolute value of the gradient across the integer ranks. Under the current policy, the slope is 204. We create a flat slope scenario by setting the slope at 22, and another steep slope scenario with the slope at 470. We also include the scenario of equal allocation of points across all rankings, in which the slope will be completely flat and both measures of curvature will be 0. To make sure that the results only come from the change in the slope of the allocation function, we fix the total points awarded to all teams in the three scenarios to 7 million. Figure 7.2 shows the point allocation slope of the 4 scenarios.

Figure 1 Point Allocation Slopes



The competitive pressure is higher in the scenarios with a steeper slope. This is because the allocation concentrates on a few high-ranked teams at the expense of teams at lower ranks. In Panel (A) of Figure 7.2, stepper slopes have higher concentration of points for top ranks, while in Panel (B), stepper slopes have less points allocated to lower ranks. When points are equally allocated, there will be no competition (except for the monetary reward). Since the curvature of the point allocation slopes represents the extent of competitive pressure for points, we denote the 4 scenarios by competitiveness for points, i.e., the completely flat slope as no competition, the flatter slope as mild competition, the original slope as original competition, and the steep slope as acute competition.

Finally, to compare the counterfactuals in Section 7.1, we also consider the three information scenarios.

Table 14 reports the results of the four scenarios. The benchmark scenario is no competition. When the competitive pressure rises, the probability of collaboration among high-ability participants will increase, as shown in the first two rows in each panel. With no information, for example, the percentage of multi-member teams has increased from 1% in the no competition scenario to 68%. These are big increases, suggesting that competition significantly affects the incentive of collaboration. When compared with the results under the original competition scenario, as reported in Table 13, the increase in collaborations between high-ability participants however is more moderate.

The last two rows in each panel in Table 14 report the mean and maximum performance in each scenario. Both performance metrics are higher in the acute competition scenario. This is because the competitive pressure for Kaggle points has increased the probability of collaborations, especially among high-ability participants. These results suggest the importance of maintaining the competitive pressure through the point allocation for creating value for sponsoring businesses. Compared with the original point allocation policy results that are reported in Table 13, the last panel of Table 14 however suggests that further increase the competitive pressure does not significantly impact the maximum team performance.

For individual participants, the average utility drops from \$2,916-3,106 in the mild competition scenario to \$929-1,055 in the acute competition scenario, more than a 60% decrease. However, the expected maximum utility increases with competitive pressure, from \$15,604-16,816 in no competition scenario to \$65,810-73,028 in acute competition scenario. The results indicate that increasing the competitive pressure will hurt an average participant's welfare but improve the top performer's welfare. Therefore, such a policy change may negatively impact the incentive of participating in competitions for average participants; however, it will boost the incentive for top participants.

To summarize the results from the two counterfactual exercises, we show how improving the informativeness of the tier status helps create a win-win situation for both sponsoring businesses and individual participants. Therefore, Kaggle should focus more on such improvement. Increasing the competitive pressure by manipulating the point allocation system will boost collaboration among participants, and increase the best performance of all teams. These will benefit sponsoring

Table 14 Performance Measures under Different Point Allocation Policies

Point Allocation	Measure	Information Structure		
		No Information	Low Information	High Information
No Competition	Percentage of Multi-Member Teams	1.0%	4.6%	5.4%
	Percentage of High-High Collaborations	0.6%	1.7%	2.4%
	Mean Utility	3106	2908	2916
	Max Utility	16816	15968	15604
	Mean Performance	-1.21	-1.04	-1.03
	Max Performance	3.26	3.32	3.34
Mild Competition	Percentage of Multi-Member Teams	5.7%	8.6%	9.8%
	Percentage of High-High Collaborations	2.1%	2.6%	2.9%
	Mean Utility	2751	2833	2861
	Max Utility	33418	44855	48951
	Mean Performance	-0.88	-0.809	-0.763
	Max Performance	3.28	3.33	3.35
Acute Competition	Percentage of Multi-Member Teams	68%	75%	76%
	Percentage of High-High Collaborations	6.8%	11.9%	13.5%
	Mean Utility	929	1039	1055
	Max Utility	65810	68819	73029
	Mean Performance	0.61	0.70	0.71
	Max Performance	3.39	3.44	3.46

businesses. An average participant's payoff will decrease while the top performer's will increase. If the priority of the platform is to attract more participants, it should avoid excessive competition for points. If the priority is the participation of top talents and sponsoring businesses, a more competitive point allocation system is preferred.

8. Conclusions

Collaboration is a common phenomenon within firms and across markets. Two main issues that are important to the efficiency of collaborations have not been fully addressed 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 use counterfactuals to show that, when the public signals (i.e., tier status) for abilities are more informative, the incentive for collaboration and the performance of collaboration will both increase. Second, individuals collaborate to compete against other collaborations. Our model incorporates competition in the payoff function. Counterfactual results show that the competitive pressure in Kaggle's competition will boost collaboration and improve team performance.

This paper makes both methodological and substantive contributions. Methodologically, we advance the literature by providing a general framework in modeling large scale one-to-one matching game that involves numerous participants. By specifying the bilateral matching decision as individuals' optimal decision, while imposing the rational expectation and market clearing constraints, our model captures a complicated market environment where incomplete information is prevalent, spillover from matching exists, and transfers between collaborators are unobserved. Substantively, we use counterfactuals to show that increasing the informativeness of signals for participants' true ability and the competitive pressure in a rank-order tournament will improve the incentive for collaborations, especially among high ability participants. This will benefit the team performance. More informative signal will also increase the payoff for participants. High competitive pressure, however, will benefit top participants but harm the payoff of general participants. Whether a crowdsourcing platform such as Kaggle should use a steeper point allocation system thus depends on what is its priority.

We have made a few simplifying assumptions to keep the model tractable. Future research should 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 and Ghili 2018) offers an alternative way to model such type of collaborations. It also models transfer or price through a Nash-in-Nash bargaining framework. Second, our model treats entry of participants as exogenous. Future research can study how the entry decision may affect collaborations, if more granular data such as click streams are readily available. Last but not least, future research should further investigate how important economic factors, such as the economy of scale, complementarity of skills, and moral hazard, separately affect the incentive and outcomes of collaborations. Currently they are combined in a reduced-form way in our model.

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.

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.

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. 2015. Cupid's invisible hand: Social surplus and identification in matching models .

Ghili, Soheil. 2018. Network formation and bargaining in vertical markets: The case of narrow networks in health insurance .

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.

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 .

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.

Appendix

A. Calculation of Team Performance Rank

We now derive the probability of team performance rank for a team $Pr(Z_{\langle i,j\rangle}(Y_{\mathcal{M}}|A_i,A_j))$.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', $Pr(a|r,r')$ 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 (7) in section 4.3. The team structure $\mathcal{M}$ could then be characterized by the propotions of team type $Pr_T(t)$ for all $t\in T$.

In section 4.2, we assume the performance of a team with type $t=\langle a,a'\rangle$, $Y(t)$ 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 component 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 p_{th} order statistic is from a particular team type t conditional on the p_{th} order statistics of Y equals y, the value of $Pr_T(p,t|Y_{(p)}=y)$ could be derive 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 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 propotion of team type t, i.e.,

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

The challenge to calculate the 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. 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, F and f are culmulative distribution and density function of Y defined in equations (A.2) and (A.3). We could simulate values from the 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 stand 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 seperately 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)

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

B. Proof: Existence of Equilibrium

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

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

Suppose the set of τ is unbounded, $\exists r, r'$, s.t. $\tau(r, r') = +\infty$, $\tau(r', r) = -\infty$. $\forall a, a'$, $Pr_{\mathcal{M}}(a, r, r'|\tau) = 1$, $Pr_{\mathcal{M}}(a', r', r|\tau) = 0$. $\forall Pr(A|R)$, $Pr_{\mathcal{M}}(r, r'|\tau) = 1$, $Pr_{\mathcal{M}}(r', r) = 0$, market equilibrium constraint is not satisfied. Thus in equilibrium the set of τ is bounded and there exists a finite number B, s.t. each coordinate of τ is in the finite interval $[-B, B]$. The set of $Pr_{\mathcal{M}}$ is bounded and closed because each coordinate of $Pr_{\mathcal{M}}$ is a probability that lies in the unit interval of $[0, 1]$. We restrict $\mathcal{D}$, the domain of $\mathcal{H}$ to be a closed and bounded subset of Euclidean space. Because each coordinate of $(Pr_{\mathcal{M}}, \tau)$ 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 $\mathcal{H}$ is continuous, and thus $\mathcal{H}$ is continuous. $\forall (Pr_{\mathcal{M}}, \tau) \in \mathcal{D}$, $\mathcal{H}(Pr_{\mathcal{M}}, \tau) \in \mathcal{D}$, because h_1 yields a mapping from a set of probabilities on to itself and h_2 comes from the market equilibrium constraint that controls the boundary of τ. Thus $\mathcal{H}$ is a continuous function from a compact and convex set $\mathcal{D}$ onto itself.

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