Dwelling on the Negative: Incentivizing Effort in Peer Prediction

Jens Witkowski
Albert-Ludwigs-Universität
Freiburg, Germany
witkowsk@informatik.uni-freiburg.de

Yoram Bachrach
Microsoft Research
Cambridge, United Kingdom
yobach@microsoft.com

Peter Key
Microsoft Research
Cambridge, United Kingdom
peter.key@microsoft.com

David C. Parkes
Harvard University
Cambridge, MA, USA
parkes@eecs.harvard.edu

Abstract

Agents are asked to rank two objects in a setting where effort is costly and agents differ in quality (which is the probability that they can identify the correct, ground truth, ranking). We study simple output-agreement mechanisms that pay an agent in the case she agrees with the report of another, and potentially penalizes for disagreement through a negative payment. Assuming access to a quality oracle, able to determine whether an agent's quality is above a given threshold, we design a payment scheme that aligns incentives so that agents whose quality is above this threshold participate and invest effort. Precluding negative payments leads the expected cost of this *quality-oracle mechanism* to increase by a factor of 2 to 5 relative to allowing both positive and negative payments. Dropping the assumption about access to a quality oracle, we further show that negative payments can be used to make agents with quality lower than the quality threshold choose to not to participate, while those above continue to participate and invest effort. Through the appropriate choice of payments, any design threshold can be achieved. This *self-selection mechanism* has the same expected cost as the cost-minimal quality-oracle mechanism, and thus when using the self-selection mechanism, perfect screening comes for free.

Introduction

We study simple output-agreement mechanisms in a setting in which agents are presented with two items that they are asked to rank. An example of such a setting are human relevance judgments for search engine results. Here, users are presented two websites together with a search query and are asked to report which of those two is better-suited for the given query. This is used to determine whether a change in the search engine's ranking algorithm improved the quality of the results and, if so, by how much. In another example, the items are suggestions what the New York City Mayor's Office could do to make New York a greener city.[1] In this wiki survey (Salganik and Levy 2012), people were presented two different suggestions and were asked to vote which of the two they found more convincing.

[1] www.allourideas.org

It is important to point out the incentive problems in these settings. In the latter, the truthfulness of reports may be a concern. A New York shop owner, for example, might agree that it makes the city cleaner having to charge customers 25 cents for each plastic bag they use, but she may also have a vested interest not to vote for this cause because it is bad for business. The primary incentive issue, however, is encouraging participants to invest costly effort in order to obtain information in the first place. In the New York setting, for example, people need to understand the issue first and then form an opinion about the presented options. In the human relevance judgments setting, it takes effort clicking on the websites, having a closer look at each, and then weighing which of the two is better suited for a given query. Faced with a payment scheme that does not address this issue properly (such as a fixed-price scheme that pays a constant amount per reported ranking), workers would maximize their hourly wage not by investing effort but by randomly clicking through each task quickly.

State-of-the-art peer prediction mechanisms do not address costly effort properly. The original peer prediction method by Miller, Resnick and Zeckhauser (2005) can incorporate costly effort but this proper scaling of payments relies on the assumption that the agents' belief model is common knowledge, which which seems unreasonable in practice. More recent mechanisms (Prelec 2004; Jurca and Faltings 2006; 2007; 2009; Witkowski and Parkes 2012a; 2012b; Radanovic and Faltings 2013) relax the assumption that the designer knows agent belief models, but lose through these relaxations the ability to know how to appropriately scale payment incentives. Moreover, none of these mechanisms support settings in which agents differ in their abilities. Recall the earlier example of the use of a wiki survey. Depending on the choices, a user can be either very knowledgeable or clueless as to which will best improve New York City's environment. The same holds true for human relevance judgments to improve search engine quality: depending on the query that is presented, a worker may or may not know how to evaluate a website based on that query.

Most closely related to the present paper is the work by Dasgupta and Ghosh (2013). Their paper is situated in a very similar model for information elicitation with unknown

ground truth where agents have individual qualities, defined just as in our model. While even simple output-agreement mechanisms induce a truthful equilibrium in this model, their key contribution is to develop a technique for eliminating other, unwanted equilibria. In a brief treatment, they comment that costly effort could be incentivized by scaling payments and that the qualities could be obtained by prescreening; e.g., through qualification tests.

In this paper, we first formalize this idea, and then use the resulting mechanism as a baseline with which to compare our mechanism, which is shown to have significantly lower cost and does not require the ability to screen participants. Rather, participants will self-select into the mechanism according to their quality. It is beyond the scope of this paper how to optimally choose the implicit quality threshold, but we note here that this trades off quality, cost and number of reports: a higher threshold results in higher expected quality of reports and lower expected cost per report but also more agents passing and thus fewer reports.

The key property of an effort-incentivizing mechanism is that the expected payment of an agent who has invested effort to observe an informative signal is higher than the expected payment of an uninformed (guessing) agent by at least the cost of effort. This property is difficult to ensure in peer prediction settings where an agent's prior beliefs are unknown to the mechanism because the expectation is formed with those unknown beliefs. Without any lower bounds on the extent of an agent's belief change from uninformed to informed, there always exist possible prior beliefs for which agents are not properly incentivized to invest effort. When such an uninformed agent is used as peer for another agent, this leads to a second-order problem because that other agent then no longer has an incentive to invest effort either.

We present two approaches to avoid this unraveling of incentives due to costly effort. We initially assume that the mechanism has access to an external quality oracle, and can preclude an agent from participating if her quality is below some design threshold. The second approach does not rely on this assumption. It avoids the aforementioned unraveling by explicitly modelling that an agent may choose not to participate in the mechanism. This "pass" action provides an outside option with utility zero and the mechanism's (potentially negative) payments are then designed such that uninformed agents or agents whose qualities are too low are better off passing than guessing. An agent whose quality is high enough, however, is better off investing effort knowing that her peer will also be informed and of high quality in equilibrium because only those participate in the first place.

The remainder of the paper is organized as follows. In the first part, we develop a baseline through an assumption of access to a quality oracle. We design payments that incentivize every agent with high enough quality to invest effort, and compare the expected cost of the mechanism that is constrained to use only non-negative payments to an unconstrained mechanism. For a uniform distribution on quality, we find that allowing for negative payments results in an expected cost 2 to 5 times lower than in the case where one can use only non-negative payments, with the cost improvement depending on the chosen quality threshold. In the second part, we then drop the assumption of access to a quality oracle, and show how to induce agents to self-select based on quality. This self-selection mechanism has the same expected cost as the quality-oracle mechanism that uses negative payments and thus admits the same, favorable cost ratio. In other words: when using our mechanism, perfect screening of agents comes for free.

The Basic Model

There are two items A and B with true order $A \succ B$; a situation that we also describe as "A is best." Each agent in a sequence of agents is presented with the items in a random order. From the perspective of a given agent i, we denote the items A_i and B_i. For example, imagine that item A_i is the item presented on the left and B_i the item presented on the right, and that the decision which item, A or B, is presented on the left side, is made uniformly at random. Due to the random bijection from items $\{A, B\}$ to an agent's subjective labels $\{A_i, B_i\}$, the prior belief of agent i is that $\Pr(A_i \succ B_i) = 0.5$.

Agent i has the option of investing effort to observe a noisy signal $\succ_i$ about the true order of the items. In particular, agent i has a *quality* q_i, which is drawn uniformly on $[0.5, 1]$. The distribution on quality is common knowledge, but each agent's quality is private. The cost of effort $C > 0$ is assumed to be identical for every agent, and also common knowledge. If agent i invests effort, the signal she receives is the true order with probability q_i; otherwise she receives the wrong order. In our analysis, it is convenient to transform quality $q_i \in [0.5, 1]$ to *normalized quality* $x_i \in [0, 1]$, so that $q_i = (1 + x_i)/2$ and x_i uniform on $[0, 1]$.

We study mechanisms that match each agent i with another agent j. Agent j is said to be the *peer* of agent i. For example, agent j can be the agent following agent i in the sequence or agent j can be chosen randomly. Agent i can also be agent j's peer but this need not be the case. We study output agreement mechanisms for which agent i receives payment $\tau_a > C > 0$ if her report agrees with that of peer agent j and $\tau_d < \tau_a$ otherwise. The main focus is to consider the impact of allowing $\tau_d < 0$. The payments in the case of agreement and disagreement are common knowledge.

An agent first decides whether to "participate" or "pass," given knowledge of her quality q_i. If she passes, she receives zero payment. If she participates, her strategic choice is whether to invest effort or not, and then to report $A_i \succ'_i B_i$ or $B_i \succ'_i A_i$ to the mechanism. The report $\succ'_i$ of agent i is mapped by the mechanism to a claim that item A is best or that item B is best, so that when we say that the report is $A \succ'_i B$ this should be understood to mean that the agent's report on A_i and B_i was mapped to mean $A \succ'_i B$. Only a participating agent can be chosen to be the peer of agent i.

If only one agent participates, we define the expected payment for this agent to be the payment she would obtain if matched against a peer who guesses; i.e., her expected payment is $(\tau_a + \tau_d)/2$.

To allow for negative payments in practice, we can imagine that the broader context requires holding at least $-\tau_d >$

0 as collateral for every agent who is interested in participating; i.e. to ask for this collateral payment upfront.

Single-Agent Perspective

In this section, we analyze the set of possible best responses of an agent. We will need this in later sections when we analyze the equilibria of our mechanisms. We also refer to Figure 1 for a graphical illustration of the game from agent i's perspective.

Agent not Investing Effort

Consider an agent who chooses to participate but not invest effort. Because items A_i and B_i are in a random bijection to A and B, the random mapping will, independent the agent's report, result in reports $A \succ_i' B$ and $B \succ_i' A$ with equal probability from the agent's perspective. For example, no matter if the agent always reports $A_i \succ_i' B_i$, always reports $B_i \succ_i' A_i$, or if she reports $A_i \succ_i' B_i$ with some probability, the agent's belief about the effect of these reports is that it is equally likely to be $A \succ_i' B$ or $B \succ_i' A$. In the same way, the effect is that an agent who does not invest effort will think it is equally likely that peer agent j's report will correspond to $A_i \succ_j' B_i$ or $B_i \succ_j' A_i$ (with respect to agent i's item space). That is, the belief of agent i in regard to the report of her peer is $\Pr(A_i \succ_j' B_i) = 0.5$. For this reason, any uninformed reporting strategy comes down to guessing uniformly with agent i receiving expected utility

$$U_i(\text{guess}) = \frac{\tau_a + \tau_d}{2} \tag{1}$$

for any report of her peer j.

Recall that the utility from not participating (i.e. from *passing*) is assumed to be zero:

$$U_i(\text{pass}) = 0. \tag{2}$$

Lemma 1 describes the primary effect of the mechanism parameters on the agents' equilibrium play:

Lemma 1. *Whether passing dominates guessing depends on payments τ_a and τ_d:*

1. *If $\tau_d > -\tau_a$, then $U_i(guess) > U_i(pass)$.*
2. *If $\tau_d < -\tau_a$, then $U_i(pass) > U_i(guess)$.*

Proof. $U_i(\text{guess}) = \frac{\tau_a+\tau_d}{2} > 0 = U_i(\text{pass}) \Leftrightarrow \tau_d > -\tau_a$ (second case analogous). □

For $\tau_d = -\tau_a$, agents are indifferent between passing and guessing. Whether passing dominates guessing or vice versa is one major difference between the two mechanisms analyzed later in this paper.

Agent Investing Effort

First observe that Lemma 1 still holds after investing effort, but that investing effort followed by guessing or passing cannot be part of any equilibrium because an agent can only incur cost and thus lose utility through investing effort followed by guess or pass.

Having invested effort, an agent can now follow more informed reporting strategies. The *truth* strategy reports the true signal received. The *counter* strategy reports the opposite order to the one received. Many other reporting strategies are available. For example, agent i can report $A_i \succ_i' B_i$ with probability $p > 0.5$ if her signal is $A_i \succ_i B_i$ and report $B_i \succ_i' A_i$ otherwise. Lemma 2 says that these other reporting strategies following a decision to invest effort are not part of any equilibrium:

Lemma 2. *If investing effort is part of a best response for an agent, then the reporting strategies "truth" or "counter" strictly dominate all other strategies.*

Proof. Since $C > 0$, and given that investing effort is part of a best response, the expected utility from investing effort must be higher than guessing. Therefore, the probability of agreement conditioned on at least one signal must be greater than 0.5. Before investing effort, the agent's subjective belief on j's report is $\Pr(A_i \succ_j' B_i) = 0.5$. Now suppose that this belief does not change after observing $A_i \succ_i B_i$, i.e. $\Pr(A_i \succ_j' B_i | A_i \succ_i B_i) = 0.5$. This would then mean that $\Pr(A_i \succ_j' B_i | B_i \succ_i A_i) = 0.5$ as well since

$$\begin{aligned} &\Pr(A_i \succ_j' B_i | A_i \succ_i B_i) \cdot \Pr(A_i \succ_i B_i) \\ &+ \Pr(A_i \succ_j' B_i | B_i \succ_i A_i) \cdot \Pr(B_i \succ_i A_i) \\ &= \Pr(A_i \succ_j' B_i) = 0.5. \end{aligned} \tag{3}$$

But then agent i's subjective belief about the probability of agreement remains unchanged and we have a contradiction. Therefore, we must have $\Pr(A_i \succ_j' B_i | A_i \succ_i B_i) \neq 0.5$. Suppose $\Pr(A_i \succ_j' B_i | A_i \succ_i B_i) > 0.5$ and so $\Pr(A_i \succ_j' B_i | B_i \succ_i A_i) < 0.5$ (follows from Equation 3). Because of this, given signal $A_i \succ_i B_i$, the agent's unique best response is to report $A_i \succ_i' B_i$ and given signal $B_i \succ_i A_i$ her unique best response is to report $B_i \succ_i' A_i$. In each case, this "truth" strategy dominates any other strategy including a mixed strategy. Similarly, if $\Pr(A_i \succ_j' B_i | A_i \succ_i B_i) < 0.5$ and so $\Pr(A_i \succ_j' B_i | B_i \succ_i A_i) > 0.5$, then the "counter" strategy dominates any other strategy. □

Given Lemma 2, it is helpful to define the *action* a_i to succinctly represent all potential equilibrium play of an agent who chooses to participate. This action is defined as follows:

$$a_i = \begin{cases} x_i & \text{, if invest effort and report truth} \\ -x_i & \text{, if invest effort and report counter} \\ 0 & \text{, if no effort, and guess.} \end{cases} \tag{4}$$

Suppose, for example, that both agent i and her peer agent invest effort and report truthfully. Agent i's expected utility,

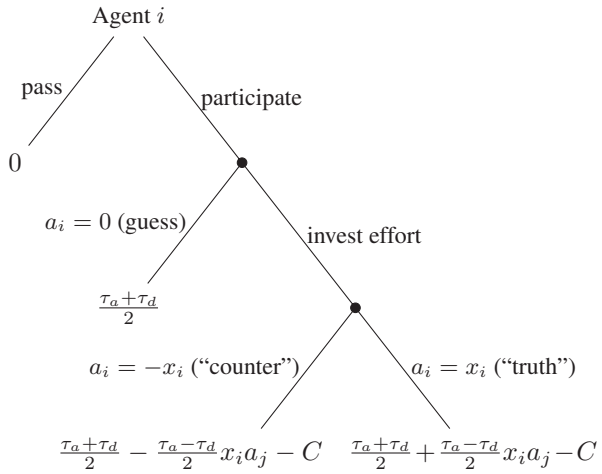

Figure 1: An illustration of agent i's decisions within the game. Note that agent i is always matched to a participating agent j.

given normalized qualities x_i and x_j, is:

$$\begin{aligned} U_i(a_i = x_i, a_j = x_j) = & \left(\frac{1+x_i}{2}\right)\left(\frac{1+x_j}{2}\right)\tau_a \\ & + \left(\frac{1-x_i}{2}\right)\left(\frac{1-x_j}{2}\right)\tau_a \\ & + \left(\frac{1+x_i}{2}\right)\left(\frac{1-x_j}{2}\right)\tau_d \\ & + \left(\frac{1-x_i}{2}\right)\left(\frac{1+x_j}{2}\right)\tau_d - C \\ = & \frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2}x_i x_j - C, \end{aligned} \tag{5}$$

where the first line is from both agreeing on the correct order, the second line is from both agreeing on the incorrect order, the third line is from agent i being correct but agent j being wrong, and the fourth line from agent i being wrong and agent j being correct.

On the other hand, if both invest effort but agent i plays truth and her peer plays counter, then τ_a and τ_d are simply exchanged, and:

$$U_i(a_i = x_i, a_j = -x_j) = \frac{\tau_a+\tau_d}{2} - \frac{(\tau_a-\tau_d)}{2}x_i x_j - C. \tag{6}$$

Suppose both were to invest effort and play counter. In this case, we again have:

$$U_i(a_i = -x_i, a_j = -x_j) = \frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2}x_i x_j - C. \tag{7}$$

Note also that if agent i invests effort, while her peer agent guesses, then her expected utility is just,

$$U_i(a_i = x_i, a_j = 0) = \frac{\tau_a+\tau_d}{2} - C. \tag{8}$$

Combining Equations 5–8, we then have the following lemma:

Lemma 3. *The expected utility for a participating agent with normalized quality x_i who takes action $a_i \in \{-x_i, +x_i\}$ is:*

$$U_i(a_i) = \frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2} \cdot a_i \cdot E[a_j \mid s_j] - C, \tag{9}$$

where $E[a_j \mid s_j]$ is the expected value of the action of peer agent j and where s_j denotes her strategy.

The expectation is taken with respect to the distribution on qualities of agents, any mixing in agent strategies, and the random process that defines an agent's signal. Following effort by agent i, her strategic interaction with the other agents is precisely captured through $E[a_j \mid s_j]$, and Equation 9 fully captures agent i's expected utility following effort.

Quality-Oracle Mechanism

In this section, we analyze mechanisms with $\tau_d > -\tau_a$ and access to a quality oracle. In the next section, we will then see that the right choice of $\tau_d < -\tau_a$ induces agents to self-select according to quality so that the mechanism no longer needs a quality oracle. The distinction between the cases $\tau_d > -\tau_a$ and $\tau_d < -\tau_a$ comes from Lemma 1. For $\tau_d > -\tau_a$, one needs some form of external quality screening because guessing strictly dominates passing, so that, without screening, every agent would participate and report something with low-quality agents guessing instead of investing effort and reporting truthfully. Knowing that some peer agents will guess, higher-quality agents would then also guess, leading to an unraveling of incentives and noise in the reported signals.

Definition 1. *For any qualification threshold $x^* > 0$, a mechanism with access to a* quality oracle *knows for every agent i whether normalized quality $x_i \geq x^*$ or $x_i < x^*$.*

In the remainder of this section, we assume the mechanism has access to a quality oracle, and uses this to only allow an agent to participate if her (normalized) quality is $x_i \geq x^* > 0$, for some threshold x^*.

When ground truth data is available for some item pairings, qualification tests can be used as an approximation for a quality oracle. Qualification tests ask every agent for reports about k item pairings for which the mechanism knows ground truth. Based on this, only those agents who agree with the ground truth on at least a fraction $x^* > 0$ of their reports are allowed to participate. Of course, such a qualification test provides only an approximate quality oracle. With $k = 10$ trials and a qualification threshold of $x^* = 0.6$ (corresponding to $q^* = 0.8$), for example, a qualification test allows an agent with $x_i = 0.2$ (corresponding to $q_i = 0.6$) to participate with 16.73% probability despite $x_i < x^*$. Similarly, an agent with $x_i = 0.7$ (corresponding to $q_i = 0.85$) misses the qualification bar with 17.98% probability. Due to the law of large numbers, these mis-classifications disappear for $k \to \infty$, so that a test with $k \to \infty$ trials approaches the behavior of an oracle.

An Effort-Inducing Equilibrium in the Quality-Oracle Mechanism

Theorem 4. *With $\tau_a - \tau_d > \frac{4C}{(x^*)^2+x^*}$ and $\tau_d > -\tau_a$, the mechanism with access to a quality oracle induces a strict Bayes-Nash equilibrium where every agent allowed to participate chooses to participate, invests effort, and reports truthfully.*

Proof. Consider an agent i who is allowed to participate after the mechanism asked the quality oracle, and assume all other agents who are allowed to participate invest effort and report truthfully. It is then a necessary and sufficient condition for the statement to be true that agent i's unique best response is to also invest effort and report truthfully. We use Equation 9 for which we first need to compute $E[a_j \mid s_j]$:

$$E[a_j|s_j] = E[x_j|x_j \geq x^*] = \frac{1}{1-x^*}\int_{x_j=x^*}^{1} x_j \,\mathrm{d}x_j = \frac{1+x^*}{2}$$

Observe that the quality distribution for peer agent j is now uniform on $[x^*, 1]$.

Since $\tau_d > -\tau_a$, we know from Lemma 1 that passing is strictly dominated by guessing, so that in every equilibrium agent i is always participating. From Lemma 2 we know that only $a_i \in \{-x_i, 0, +x_i\}$ can be best responses of an agent that participates. Now since $E[a_j|s_j] = \frac{1+x^*}{2} > 0$ and $x_i \geq x^* > 0$, by Equation 9, $a_i = -x_i$ cannot be part of a best response either. It then remains to determine the values τ_a, τ_d with $\tau_d > -\tau_a$ for which an agent i with quality $x_i = x^*$ is better off playing $a_i = x_i$ than $a_i = 0$ by setting $U_i(a_i = x^*) > U_i(a_i = 0)$:

$$\frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2}\cdot x^* \cdot E[a_j|s_j] - C > \frac{\tau_a+\tau_d}{2}$$
$$\Leftrightarrow \frac{\tau_a-\tau_d}{2}\cdot\frac{x^*(1+x^*)}{2} > C \Leftrightarrow \tau_a - \tau_d > \frac{4C}{(x^*)^2+x^*}.$$

This completes the proof. □

As always in peer prediction, this equilibrium is not unique (e. g., Waggoner and Chen, 2013). In particular, all agents guessing is also an equilibrium.

Expected Cost of Quality-Oracle Mechanism

Now that we know the constraint on τ_a and τ_d such that for a given quality threshold x^*, there is a Bayes-Nash equilibrium where all agents with quality higher than the threshold invest effort and are truthful, how should payments τ_a and τ_d be set to minimize the expected cost given C and x^*?

For each agent who participates, the expected cost in the truth equilibrium of the quality-oracle mechanism is given by

$$\begin{aligned} &E[\text{cost for participating agent}] \\ =&\frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2}\cdot\frac{x^*+1}{2}\cdot\frac{x^*+1}{2} \\ =&\frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2}\cdot\frac{(x^*+1)^2}{4} \\ =&\frac{1}{2}(\tau_a+\tau_d) + \frac{(x^*+1)^2}{8}(\tau_a-\tau_d). \end{aligned} \tag{10}$$

Given this, the optimization problem to find the cost-minimizing mechanism parameters becomes:

$$\begin{aligned} \text{minimize}\quad & \frac{1}{2}(\tau_a+\tau_d) + \frac{(x^*+1)^2}{8}(\tau_a-\tau_d) \\ \text{s.t.}\quad & \tau_a - \tau_d > \frac{4C}{(x^*)^2+x^*} \\ & \tau_a + \tau_d > 0 \end{aligned} \tag{11}$$

The remainder of the section is organized as follows. We first solve this optimization problem, and then impose the additional requirement of non-negative payments, i.e. $\tau_d \geq 0$. Having done this, we quantify how much more the mechanism has to pay in expectation because of this restriction.

Allowing for Negative Payments. We solve the optimization problem as given in (11) using a variable change. Let $\tau_a = \tau + \Delta$ and $\tau_d = \tau - \Delta$ for new variables τ and Δ. Substituting into (11) and solving the optimization problem immediately gives $\Delta = 2C/\left((x^*)^2+x^*\right) + \epsilon$, and $\tau = \epsilon$, with $\epsilon > 0$ and $\epsilon \to 0$. Substituting again for τ_a and τ_d, we obtain the cost-minimizing mechanism parameters:

$$\tau_a = \frac{2C}{(x^*)^2+x^*} + \epsilon \tag{12}$$

and

$$\tau_d = -\frac{2C}{(x^*)^2+x^*}. \tag{13}$$

Given this, the expected cost to the mechanism for each agent who chooses to participate is

$$\begin{aligned} &E[\text{minimal cost for participating agent}|\tau_d > -\tau_a] \\ =&\frac{1}{2}(\tau_a+\tau_d) + \frac{(x^*+1)^2}{8}(\tau_a-\tau_d) \\ =&\frac{1}{2}(\epsilon) + \frac{(x^*+1)^2}{8}\left(\frac{4C}{(x^*)^2+x^*}+\epsilon\right), \end{aligned} \tag{14}$$

and for $\epsilon \to 0$, we have:

$$\begin{aligned} &\lim_{\epsilon\to 0} E[\text{minimal cost for participating agent}|\tau_d > -\tau_a] \\ =&\frac{(x^*+1)^2}{8}\cdot\frac{4C}{(x^*)^2+x^*} = \frac{(x^*)^2+2x^*+1}{2(x^*)^2+2x^*}\,C \\ =&\left(\frac{1}{2x^*}+\frac{1}{2}\right)\,C. \end{aligned} \tag{15}$$

Requiring Non-negative Payments. Let's now suppose that we seek to minimize the expected cost of the quality-oracle mechanism subject to $\tau_d \geq 0$. The second constraint in (11) is always satisfied with $\tau_d \geq 0$. We now argue that it must be that $\tau_d = 0$ in an optimal solution. Assume this was not the case, so that $\tau_d = y$ for some $y > 0$. Then the left-hand side of the first constraint can be kept at the same level by setting $\tau_d := 0$ and $\tau_a := \tau_a - y$, which would lower the first part of the objective function and leave the second

part unchanged. So for minimal non-negative payments, we require:

$$\tau_d = 0 \tag{16}$$

After inserting $\tau_d = 0$ back into the optimization problem, we have

$$\tau_a = \frac{4C}{(x^*)^2 + x^*} + \epsilon \tag{17}$$

with $\epsilon \to 0$. Based on this, the expected cost to the mechanism for each agent who chooses to participate is

$$\begin{aligned} &E[\text{minimal cost for participating agent}|\tau_d \geq 0] \\ =&\frac{1}{2}(\tau_a + \tau_d) + \frac{(x^*+1)^2}{8}(\tau_a - \tau_d) \\ =&\frac{(x^*)^2 + 2x^* + 5}{8}\left(\frac{4C}{(x^*)^2 + x^*} + \epsilon\right) \end{aligned} \tag{18}$$

and for $\epsilon \to 0$, we have:

$$\begin{aligned} &\lim_{\epsilon \to 0} E[\text{minimal cost for participating agent}|\tau_d \geq 0] \\ =&\frac{(x^*)^2 + 2x^* + 5}{8}\left(\frac{4C}{(x^*)^2 + x^*} + \epsilon\right) \\ =&\frac{(x^*)^2 + 2x^* + 5}{2(x^*)^2 + 2x^*}C. \end{aligned} \tag{19}$$

Relative Cost of Requiring Non-Negative Payments. Clearly, constraining the mechanism's payments to be non-negative can only increase the expected cost (fixing the cost to agents for effort C and the design parameter x^* above which agents will choose to participate and invest effort). But how much more expensive is it when the mechanism is restricted to only use non-negative payments? Recall that in both cases we insist that agents must have an incentive to participate, i.e. the expected utility for an agent who participates remains non-negative.

Theorem 5. *Fixing quality threshold x^*, the expected cost of the cost-optimized quality-oracle mechanism increases by a factor of*

$$\frac{4}{(x^*+1)^2} + 1$$

when constraining the mechanism to non-negative payments $\tau_a, \tau_d \geq 0$. This is an increase between 2 (for $x^* \to 1$) and 5 (for $x^* \to 0$).

Proof. The result follows from dividing Equation 19 by Equation 15:

$$\begin{aligned} &\frac{\lim_{\epsilon \to 0} E[\text{minimal cost for participating agent}|\tau_d \geq 0]}{\lim_{\epsilon \to 0} E[\text{minimal cost for participating agent}|\tau_d > -\tau_a]} \\ =&\frac{\frac{(x^*)^2+2x^*+5}{2(x^*)^2+2x^*}C}{\frac{(x^*)^2+2x^*+1}{2(x^*)^2+2x^*}C} = \frac{(x^*)^2 + 2x^* + 5}{(x^*)^2 + 2x^* + 1} \\ =&\frac{(x^*+1)^2 + 4}{(x^*+1)^2} = \frac{4}{(x^*+1)^2} + 1 \end{aligned} \tag{20}$$

Since $(x^*+1)^2$ is strictly increasing in x^*, the term $4/(x^*+1)^2 + 1$ is strictly decreasing. The statement follows after inserting $x^* = 0$ and $x^* = 1$. □

Self-Selection Mechanism

In this section, we drop the assumption that the mechanism has access to a quality oracle. At the same time, we consider the effect of setting $\tau_d < -\tau_a$ so that passing dominates guessing (Lemma 1). The main result is that we identify a natural equilibrium in which agents self-select according to their quality x_i, such that every agent over quality threshold x^* invests effort and is truthful, and every agent below the threshold is passing.

There are several advantages of self selection when compared to qualification tests: first, qualification tests are quality oracles only with infinitely many samples. Second, they are wasteful because agents need to be paid for test questions to which the answer is already known. Third, self selection is more flexible than qualification tests in that it readily adapts to changes in the nature of tasks without any re-testing. In the human relevance judgments setting, for example, the quality of an agent does not have to be fixed but can depend on the given search query. Finally, self selection does not require any ground truth data.

An Effort-Inducing Equilibrium in the Self-Selection Mechanism

Let $s_i(x_i)$ denote the strategy that maps agent i's quality type x_i to her action (e.g. "pass" or "guess").

Theorem 6. *With $\tau_d < -\tau_a$, the mechanism without access to a quality oracle induces a Bayes-Nash equilibrium where every agent i plays the following strategy:*

$$s_i(x_i) = \begin{cases} \textit{invest effort and report truthfully}, & \textit{if } x_i \geq x^* \\ \textit{pass}, & \textit{if } x_i < x^* \end{cases}$$

where

$$x^* = \sqrt{\frac{9}{4} - \frac{4(\tau_a - C)}{\tau_a - \tau_d}} - \frac{1}{2}.$$

For an agent with $x_i \neq x^*$ this is a strict best response.

Proof. It is sufficient to show that given peer agent j plays $s_j(x_j)$, it is a best response of agent i to play $s_i(x_i)$, and the unique best response if $x_i \neq x^*$. Inserting $s_j(x_j)$ into Equation 9, $E[a_j \mid s_j]$ is identical to its value in the quality-oracle mechanism:

$$E[a_j|s_j] = E[x_j|x_j \geq x^*] = \frac{1}{1-x^*}\int_{x_j=x^*}^{1} x_j \, \mathrm{d}x_j = \frac{1+x^*}{2}.$$

By Lemma 1, we know that passing strictly dominates guessing, so in order to determine the indifference point between passing and investing effort followed by truthful reporting, we set $U_i(a_i = x^*) = U_i(\text{pass}) = 0$ and obtain

$$\begin{aligned} &\frac{\tau_a + \tau_d}{2} + \frac{\tau_a - \tau_d}{2} \cdot x^* \cdot E[a_j|s_j] - C = 0 \\ \Leftrightarrow\ &\frac{\tau_a + \tau_d}{2} + \frac{(\tau_a - \tau_d)(x^*(1+x^*))}{4} - C = 0 \\ &\Leftrightarrow x^* + (x^*)^2 = 2 - \frac{4(\tau_a - C)}{\tau_a - \tau_d} \\ &\Leftrightarrow x^* = \pm\sqrt{\frac{9}{4} - \frac{4(\tau_a - C)}{\tau_a - \tau_d}} - \frac{1}{2}. \end{aligned} \tag{21}$$

First observe that $\frac{4(\tau_a - C)}{\tau_a - \tau_d} > 0$ because $\tau_a > C$. Now, $\tau_a - C < \tau_a$ and $\tau_a - \tau_d > 2\tau_a$ since $-\tau_d > \tau_a$. Therefore, it holds that $4(\tau_a - C)/(\tau_a - \tau_d) < 4\tau_a/2\tau_a = 2$ and so only the positive solution of the square root is within the $[0, 1]$ bounds for x^*. Strictness for $x_i \neq x^*$ follows from Equation 9 strictly increasing with a_i for $E[a_j \mid s_j] > 0$. □

Expected Cost of Self-Selection Mechanism

Since agent j's equilibrium play is the same as in the previous section, so is the equation for the expected cost of the mechanism (Equation 10). For the self-selection mechanism, the difference is that the equilibrium conditions do not allow the same simple analysis as for the quality-oracle mechanism to find the cost-minimal payments because the equilibrium condition from Theorem 6 do not have the same simple structure as those from Theorem 4. We thus insert the equilibrium condition into Equation 10 and obtain (the second line in Equation 22 is derived from the second line in Equation 21):

$$
\begin{aligned}
&E[\text{cost for participating agent}] \\
=&\frac{\tau_a + \tau_d}{2} + \frac{\tau_a - \tau_d}{2} \cdot \frac{(x^*+1)^2}{4} \\
=&\frac{\tau_a + \tau_d}{2} + \frac{2C - (\tau_a + \tau_d)}{x^* + (x^*)^2} \cdot \frac{(x^*+1)^2}{4} \\
=&\frac{\tau_a + \tau_d}{2} + \frac{(x^*+1)\big(2C - (\tau_a + \tau_d)\big)}{4x^*} \\
=&\frac{\tau_a + \tau_d}{2} - \frac{(\tau_a + \tau_d)(x^*+1)}{4x^*} + \frac{2C(x^*+1)}{4x^*} \\
=&\frac{\tau_a + \tau_d}{2} - \frac{\tau_a + \tau_d}{2}\left(\frac{1}{2} + \frac{1}{2x^*}\right) + \frac{C}{2} + \frac{C}{2x^*} - C + C \\
=&\frac{\tau_a + \tau_d}{2}\left(\frac{1}{2} - \frac{1}{2x^*}\right) - C\left(\frac{1}{2} - \frac{1}{2x^*}\right) + C \\
=&\left(\frac{\tau_a + \tau_d}{2} - C\right)\left(\frac{1}{2} - \frac{1}{2x^*}\right) + C \\
=&\left(C - \frac{\tau_a + \tau_d}{2}\right)\left(\frac{1}{2x^*} - \frac{1}{2}\right) + C
\end{aligned}
\tag{22}
$$

Both factors of the first part of this equation are always positive for $\tau_d < -\tau_a$ and $x^* \in (0, 1)$. Fixing x^*, the minimal payments are thus setting $\tau_d = -\tau_a - \epsilon$ with $\epsilon > 0$ and $\epsilon \to 0$, so that $(\tau_a + \tau_d)/2 \to 0$. Setting $\tau_d = -\tau_a - \epsilon$ leaves us with one degree of freedom, and τ_a can still be used to implement any x^*, since

$$\lim_{\epsilon \to 0} \frac{4(\tau_a - C)}{\tau_a - \tau_d} = \lim_{\epsilon \to 0} \frac{4(\tau_a - C)}{2\tau_a + \epsilon} = 2 - \frac{2C}{\tau_a},$$

so that

$$x^* := \sqrt{\frac{9}{4} - \left(2 - \frac{2C}{\tau_a}\right)} - \frac{1}{2} = \sqrt{\frac{1}{4} + \frac{2C}{\tau_a}} - \frac{1}{2}, \tag{23}$$

can be set to any value between 0 and 1. Solving Equation 23 for τ_a, we thus obtain the cost-minimal payments

$$\tau_a = \frac{2C}{(x^*)^2 + x^*}$$

and

$$\tau_d = -\frac{2C}{(x^*)^2 + x^*} - \epsilon.$$

For $\epsilon \to 0$, these are identical to the cost-minimal payments of the quality-oracle mechanism with $\tau_d > -\tau_a$. Therefore, the self-selection mechanism with $\tau_d < -\tau_a$ has the same expected cost, with the added benefit of obtaining perfect screening through self-selection. Theorem 7 then follows immediately:

Theorem 7. *For fixed quality threshold x^*, the expected cost of the cost-optimized self-selection mechanism is lower than the expected cost of the cost-optimized quality-oracle mechanism constrained to non-negative payments by a factor of*

$$\frac{4}{(x^*+1)^2} + 1.$$

In light of this, we believe that practitioners should strongly consider allowing negative payments: they significantly lower the cost for effort-incentivizing peer prediction mechanisms and provide a free way to perfectly screen based on quality in equilibrium.

Conclusions

We have presented an analysis of simple output-agreement mechanisms for incentivizing effort and providing screening for worker quality. In closing, we emphasize two main points:

First, peer prediction with effort incentives is expensive if simple output agreement can only use non-negative payments. For example, with effort cost $C > 0$ and a quality threshold of $q^* = 0.8$ (i.e., designing for only the top 40% of quality in the market), the expected cost for the cost-minimal, non-negative-payment output-agreement mechanism is $3.4C$. Allowing for negative payments, the expected cost for the mechanism decreases to $1.3C$. This improvement occurs even if the designer has access to a way to perfectly screen for the quality of participants and comes about through maintaining the differential incentive for agreement over disagreement, while reducing the expected payment so that agents with quality at the threshold are just indifferent between participation and not.

In addition to lower expected cost on behalf of the mechanism, choosing negative payments in a way that they disincentivize guessing induces an equilibrium where agents self-select according to the selection criterion—in effect, perfect screening comes for free. We do not believe that this principle is restricted to selection for quality. For example, it could also be applied when all participants have the same quality but differ in cost, and agents self-select according to cost.

In markets with task competition, an agent's outside option may not be to pass and obtain utility zero but to work on another task with positive expected utility. For such a competitive setting, we conjecture that the relative cost benefit of negative payments decreases but that incentivizing agents with low quality to pass still has the benefit that it induces agents to self select according to their qualities.

Regarding the practicality of our approach, we think it is useful to separate the assumptions made for the analysis (such as the uniform prior on the agents' quality) and the simplicity (and thus practical robustness) of simple output-agreement mechanisms. In practice, a designer only needs to set two parameters, agreement payment τ_a and disagreement payment τ_d, and this could be achieved adaptively. Our main theoretical results suggest the opportunity for significant cost savings, along with screening of agents according to quality through self selection.

There are several interesting directions for future work. First, it would be interesting to evaluate our mechanisms experimentally. Second, we plan to extend our analysis to more sophisticated belief models where an agent may believe that she holds a minority opinion after investing effort. This is currently precluded by the way an agent's quality is modeled because after investing effort, an agent observes ground truth with probability larger than 50%. In particular, we intend to study effort incentives for peer-prediction mechanisms that are not just simple output agreement, such as the robust Bayesian truth serum (Witkowski and Parkes 2012a). Finally, combining machine learning models with peer prediction is an interesting direction, presenting potential applications across multiple domains. One interesting area of application is peer grading in massively open online courses (MOOCs) where students grade other students' assignments. The machine learning work by (Piech et al. 2013) learns each grader's quality and bias from Coursera[2] data with some success but ignores effort incentives for accurate grading. We believe that incorporating proper incentives for effort similar to the techniques of this paper will increase the performance of these algorithms.

Acknowledgements

Part of this work was done while Jens Witkowski was a research intern at Microsoft Research Cambridge (UK). This research is also supported in part through NSF award 0915016.

References

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *Proceedings of the 22nd ACM International World Wide Web Conference (WWW'13)*.

Jurca, R., and Faltings, B. 2006. Minimum Payments that Reward Honest Reputation Feedback. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06)*.

Jurca, R., and Faltings, B. 2007. Robust Incentive-Compatible Feedback Payments. In *Trust, Reputation and Security: Theories and Practice*, volume 4452 of *LNAI*. Springer-Verlag. 204–218.

Jurca, R., and Faltings, B. 2009. Mechanisms for Making Crowds Truthful. *Journal of Artificial Intelligence Research (JAIR)* 34:209–253.

[2] www.coursera.com

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51(9):1359–1373.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM'13)*, 153–160. International Educational Data Mining Society.

Prelec, D. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306(5695):462–466.

Radanovic, G., and Faltings, B. 2013. A Robust Bayesian Truth Serum for Non-binary Signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13)*.

Salganik, M. J., and Levy, K. E. C. 2012. Wiki surveys: Open and quantifiable social data collection. preprint, http://arxiv.org/abs/1202.0500.

Waggoner, B., and Chen, Y. 2013. Information Elicitation Sans Verification. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC'13)*.

Witkowski, J., and Parkes, D. 2012a. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*.

Witkowski, J., and Parkes, D. 2012b. Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*.