Notes on Peer Prediction Mechanisms

For an exposition on this idea, see: `https://blog.kleros.io/incentivizing-jurors-to-honestly-report-uncommon-answers-deposit-sizes-lazy-strategies-and-peer-prediction/`

In these notes, we pursue this analysis with a somewhat more nuanced model. Instead of only allowing jurors to exert a fixed amount of effort or exert no effort for a given task, we allow for them to exert a variable amount of effort. Then, the chances of a juror producing the correct result and being rewarded depend on the effort she puts in.

1 Model and notation

We take N to be the set of participants, $|N| = n$. Take $i \in N$ to be a given user.

Let $x_i \in [0, \infty)$. This will represent the amount of effort that is performed on a given task by i. We will simply write x for this amount of effort when the relevant user is clear from context.

We will continue tasks as having one of two possible difficulty levels, z_1 or z_2. The task with difficulty level z_1 occurs with probability p_1 whereas the task with difficulty level z_2 occurs with probability p_2, with $p_1 + p_2 = 1$.

The difficulty curves of our tasks will be defined by four parameters: A, B, D_1, and D_2 such that $A \in (0, 1)$, $B \leq 0$, $0 \leq D_2 \leq D_1$, and

$$\frac{1}{2} \leq (A + B)D_2 \leq (A + B)D_1.$$

Then the probability a juror has of determining the correct answer to a task of difficulty z_j when making effort x is given by:

$$f(x, z_j) = (A + Be^{-x}) \cdot D_j.$$

Thus, z_2 is essentially the harder difficulty level as, for a fixed effort, the chance of the juror selecting the correct answer is reduced as $0 \leq D_2 \leq D_1$.

We will compare two mechanisms, a mechanism that involves peer predictions and a similar default mechanism that does not.

Let $f \geq 0$, $d \geq 0$, $\lambda \geq 0$, and $\alpha \geq 0$. These values will serve as parameters in the payoff functions for the two mechanisms. Essentially, we think of f as a fee that is being paid to participants and d as a deposit that they can lose

for incorrect responses. Furthermore, λ will be a risk-aversion parameter that will be used to weight participants' losses differently than their gains. In the calculation of participant utility, losses will be multiplied by λ; so then a risk-neutral participant will have $\lambda = 1$ indicating that losses and gains are weighted equally whereas a risk-averse participant will have $\lambda > 1$. In the peer-prediction based mechanism, α will be used to weight how much the user is rewarded or penalized for the vote she provides versus how much she is rewarded or penalized for the predictions she provides for the votes of her peers.

For the default mechanism, the payoff is given by:

$$\left\{ \begin{array}{ccc} f - x & : & \text{right response} \\ f - \lambda \cdot d - x & : & \text{wrong response} \end{array} \right\}$$

Essentially, the participant is rewarded f regardless and is penalized d if she is incorrect. Furthermore, the user's payoff accounts for the effort she has exerted by the term that subtracts x. In Kleros, fees that jurors receive as payment for tasks are in ETH whereas penalties are in PNK - this use of separate currencies can justify applying the risk aversion parameter here only to deposit term, rather than trying to capture the net losses. Variations that handle this question can be considered in future work. Note that we are assuming that the mechanism has access to the "right response". For Kleros, this can be justified at least heuristically as cases can be appealed, so cases that are resolving to the "wrong response" can be reconsidered by larger and larger panels of jurors.

Now we consider the peer prediction mechanism. We denote the option that the user votes for as v and her predictions of the rates at which users will vote for v and the other option as y and $1 - y$ respectively. We take $\hat{y}$ to be the average of the rates at which the other participants vote for v.

Then user's the payoff is given by:

$$\left\{ \begin{array}{ccc} f + \alpha(2y - (y^2 + (1-y)^2) + 1) - x & : & \text{right response} \\ f - \lambda \cdot d \cdot (1 - \hat{y}) + \alpha(2(1-y) - (y^2 + (1-y)^2) + 1) - x & : & \text{wrong response} \end{array} \right\}$$

Note that as $y \in [0, 1]$, $2y - (y^2 + (1-y)^2) + 1 \geq 0$ and $2(1-y) - (y^2 + (1-y)^2) + 1 \geq 0$. Thus, we do not apply the risk-aversion factor to losses to these terms.

Finally, we will have a parameter $c \in (p_1(A+B)D_1 + p_2(A+B)D_2, p_1AD_1 + p_2AD_2)$ that will act as a target level of accuracy for our mechanisms. Namely, for the sake of being able to make a fair comparison between the default and the peer-prediction mechanism, we perform the following:

- Consider the Nash equilibria generated by the default and the peer-prediction mechanisms for given values of f, d, and α. This yields an amount of effort in equilibrium that we denote x_* for a given mechanism.
- Imagine that a system administrator or a governance process chooses the values f, d, and α such that

$$p_1 f(x_*, z_1) + p_2 f(x_*, z_2) = c$$

for each mechanism. Note that by our assumptions on c, c is in the range of $p_1 f(x, z_1) + p_2 f(x, z_2)$.

This will generally lead to different choices of f and d for the two mechanisms, but then we will explore in Section 3 how the two mechanisms, calibrated to the same level of average accuracy c, compare in terms of how risk aversion impacts the utility of participants.

In this work, we will consider all participants as risk-averse with risk aversion factor λ, but otherwise economically rational unless specified. Namely, participants attempt to maximize their risk-adjusted utility functions. In Section **??**, we will consider participants that gain utility from harming the other participants, namely that are willing to take on losses in order to reduce the payoffs of others.

2 Equilibrium behaviour

The expected utility under the default mechanism, taking into account the participant's chances of producing the correct response, is given by:

$$u(x, z_j) = f - \lambda \cdot d(1 - f(x, z_j)) - x. \tag{1}$$

To determine the expect utility of the peer-prediction mechnaism, we must consider the peer-predictions that are produced in equilibrium.

Proposition 1. *Suppose that a rational user exerts effort x for a task of difficulty z_j. Then she will provide a prediction scores of $f(x, z_j)$ and $1 - f(x, z_j)$ for the options she perceives as more likely and less likely respectively.*

Proof. This follows from

$$\left\{ \begin{array}{ccc} 2y - (y^2 + (1-y)^2) & : & \text{right response} \\ 2(1-y) - (y^2 + (1-y)^2) & : & \text{wrong response} \end{array} \right\}$$

being a proper scoring rule.

□

Then, when calculating the expected utility of i under our rationality assumptions, we can substitute $y = f(x, z_j)$ as this is the i's optimal choice. Note that $\hat{y}$ depends on the other users' choices, and is hence independent of the $x = x_i$. Thus, i has an expected utility function of:

$$\begin{aligned} u(x, z_j) = f &- \lambda \cdot d(1 - f(x, z_j)) \cdot (1 - \hat{y}) + f(x, z_j) \cdot \alpha \left[2f(x, z_j) - (f(x, z_j)^2 + (1 - f(x, z_j))^2) + 1 \right] \\ &+ (1 - f(x, z_j)) \cdot \alpha \left[2(1 - f(x, z_j)) - (f(x, z_j)^2 + (1 - f(x, z_j))^2) + 1 \right] - x \end{aligned}$$

$$= f - \lambda \cdot d(1 - f(x, z_j)) \cdot (1 - \hat{y}) + \alpha \left[f(x, z_j)^2 + (1 - f(x, z_j))^2 + 1 \right] - x. \tag{2}$$

Now we can compute the effort $x = x_i$ that i exerts in equilibrium for each mechanism.

Proposition 2. *The effort x_i exerted by i in equilibrium under the default mechanism is given as follows:*

$$x_i = \left\{ \begin{array}{ccc} -\ln\left[\frac{-1}{\lambda d B D_j}\right] & : & \lambda d \geq \frac{-1}{BD_j} \\ 0 & : & \lambda d < \frac{-1}{BD_j} \end{array} \right\}.$$

This corresponds to:

$$f(x_i, z_j) = \left\{ \begin{array}{ccc} AD_j - \frac{1}{\lambda d} & : & \lambda d \geq \frac{-1}{BD_j} \\ (A+B)D_j & : & \lambda d < \frac{-1}{BD_j} \end{array} \right\}.$$

Proof. This is a straightforward optimization argument taking the derivative of $u(x, z_j)$ given by equation 1 with respect to x.

□

Proposition 3. *Let $\alpha = 0$. There exists an equilibrium where the effort x_i exerted by i under the peer-prediction mechanism is given as follows:*

$$x_i = \left\{ \begin{array}{ccc} -\ln\left[\frac{-AD_j + \sqrt{(AD_j)^2 - \frac{4}{\lambda d}}}{2BD_j}\right] & : & \lambda d \geq \frac{-1}{(A+B)D_j^2 B} \\ 0 & : & \lambda d < \frac{-1}{(A+B)D_j^2 B} \end{array} \right\}.$$

This corresponds to

$$f(x_i, z_j) = \left\{ \begin{array}{ccc} \frac{AD_j + \sqrt{(AD_j)^2 - \frac{4}{\lambda d}}}{2} & : & \lambda d \geq \frac{-1}{(A+B)D_j^2 B} \\ (A+B)D_j & : & \lambda d < \frac{-1}{(A+B)D_j^2 B} \end{array} \right\}.$$

Proof. Noting that $\hat{y}$ is constant with respect to x_i, one can do an optimization argument taking the derivative of $u(x, z_j)$ given by equation 2 with respect to x_i to see that i's optimal effort x_i satisfies:

$$BD_j e^{-x} = \frac{1}{-\lambda d(1-\hat{y})} \tag{3}$$

as long the optimal value of x is not on a boundary point, namely as long as the corresponding $x > 0$. Rearranging equation 3, we note that $x > 0$ is equivalent to $\lambda d > \frac{-1}{(1-\hat{y})BD_j}$.

Thus, we have that the optimal value of effort x_i for participant i is such that:

$$f(x_i, z_j) = \left\{ \begin{array}{ccc} AD_j - \frac{1}{\lambda d(1-\hat{y})} & : & \lambda d \geq \frac{-1}{(1-\hat{y})BD_j} \\ (A+B)D_j & : & \lambda d < \frac{-1}{(1-\hat{y})BD_j} \end{array} \right\}. \tag{4}$$

Denote $t = f(x, z_j)$. In equilibrium, the condition of equation 4 should hold for each participant. We see that for any value of λd there is an equilibrium where all participants exert the same effort and thus:

$$f(x, z_j) = AD_j + BD_j e^{-x} = 1 - \hat{y}.$$

In the case where $t = (A + B)D_j$, then $\frac{-1}{(1-\hat{y})BD_j} = \frac{-1}{(A+B)BD_j^2}$. Then if

$$\lambda d < \frac{-1}{(A + B)BD_j^2}$$

we have an equilibrium where $t = (A + B)D_j$.

On the other hand, equilibria that satisfy the first case of equation 4 correspond to values of t that solve:

$$t = AD_j + BD_j e^{-x} = AD_j + \frac{1}{-\lambda dt}$$

Solving this equation for t shows us that there is a solution of the form:

$$f(x, z_j) = t = \frac{AD_j + \sqrt{(AD_j)^2 - \frac{4}{\lambda d}}}{2}$$

if $(AD_j)^2 - \frac{4}{\lambda d} \geq 0$, and if $(AD_j)^2 - \frac{4}{\lambda d} < 0$ there are no solutions for t.

However, as $A + B \geq \frac{1}{2}$,

$$A^2 - 2A + 1 \geq 0 \forall A \Rightarrow A^2 + 4(A + B)B \geq 0$$

$$\Rightarrow (AD_j)^2 - \frac{4}{\lambda d} \geq 0$$

for any $\lambda d \geq \frac{-1}{(A+B)BD_j}$.

Furthermore, we will see that if $\lambda d \geq \frac{-1}{(A+B)BD_j}$ and $t = \frac{AD_j+\sqrt{(AD_j)^2-\frac{4}{\lambda d}}}{2}$, then

$$\lambda d \geq \frac{-1}{tBD_j}.$$

Indeed, this is equivalent to

$$-\left(\frac{AD_j + \sqrt{(AD_j)^2} - \frac{4}{\lambda d}}{2}\right) BD_j \lambda d \geq 1.$$

However, a straightforward first derivative argument shows that the function

$$g(x) = -\left(\frac{AD_j + \sqrt{(AD_j)^2} - \frac{4}{x}}{2}\right) BD_j x$$

is monotonically increasing for all positive x where the square roots are defined. However, one notes that $g\left(\frac{-1}{(A+B)BD_j}\right) = 1$.

Thus, if $\lambda d \geq \frac{-1}{(A+B)BD_j}$, the value of $f(x_i, z_j) = t = \frac{AD_j + \sqrt{(AD_j)^2 - \frac{4}{\lambda d}}}{2}$ satisfies the first condition of equation 4.

□

Lemma 1. *Let $c < p_1 AD_1 + p_2(A+B)D_2$. Suppose*

$$\frac{p_1}{p_1 AD_1 + p_2(A+B)D_2 - c} \leq \frac{-1}{BD_2}$$

and choose d such that

$$\lambda d = \frac{p_1}{p_1 AD_1 + p_2(A+B)D_2 - c}.$$

Then $f(x_*, z_2) = (A+B)D_2$ and

$$p_1 f(x_*, z_1) + p_2 f(x_*, z_2) = c$$

for x_* the equilibrium effort in the default mechanism.

Proof. Recall $D_1 \geq D_2 \Rightarrow BD_1 \leq BD_2 \Rightarrow \frac{-1}{BD_1} \leq \frac{-1}{BD_2}$.

Take λd as above. Suppose that

$$\lambda d = \frac{p_1}{p_1 AD_1 + p_2(A+B)D_2 - c} \leq \frac{-1}{BD_1}.$$

Then

$$c \leq p_1(A+B)D_1 + p_2(A+B)D_2,$$

which violates our hypothesis on c. Thus, $\lambda d > \frac{-1}{BD_1}$.

Then, we must have $f(x_*, z_1)$ in the first case of Proposition 2 and $f(x_*, z_2)$ in the second case. Thus, indeed,

$$p_1 f(x_*, z_1) + p_2 f(x_*, z_2) = p_1 \left(AD_1 - \frac{1}{\lambda d} \right) + p_2 \left((A+B)D_2 \right) = c.$$

□

Lemma 2. *Let $\alpha = 0$ and let $c < p_1 AD_1 + p_2(A+B)D_2$. Suppose*

$$\frac{1}{\left(AD_1 - \frac{c - p_2(A+B)D_2}{p_1} \right) \cdot \left(\frac{c - p_2(A+B)D_2}{p_1} \right)} \leq \frac{-1}{(A+B)BD_2^2}$$

and choose d such that

$$\lambda d = \frac{1}{\left(AD_1 - \frac{c - p_2(A+B)D_2}{p_1} \right) \cdot \left(\frac{c - p_2(A+B)D_2}{p_1} \right)}.$$

Then $f(x_*, z_2) = (A+B)D_2$ and

$$p_1 f(x_*, z_1) + p_2 f(x_*, z_2) = c$$

for x_* the equilibrium effort in the peer-prediction mechanism.

Proof. Define the function

$$g(x) = \frac{1}{\left(AD_1 - \frac{x-p_2(A+B)D_2}{p_1}\right) \cdot \left(\frac{x-p_2(A+B)D_2}{p_1}\right)}.$$

A straightforward argument shows that $g(x)$ is monotone increasing for x such that both of the following conditions hold: $\frac{x-p_2(A+B)D_2}{p_1} \geq \frac{AD_1}{2}$ and $AD_1 - \frac{x-p_2(A+B)D_2}{p_1} \geq 0$.

Note that, as $A+B \geq \frac{1}{2}$ and $A < 1$, $A+2B > 0 \Rightarrow \frac{AD_1}{2} < (A+B)D_1$. So $g(x)$ is monotonically increasing for all $x \in (p_1(A+B)D_1 + p_2(A+B)D_2, p_1AD_1 + p_2(A+B)D_2)$. However, $x = c$ is in this interval, so

$$\frac{-1}{(A+B)BD_1^2} = g(p_1(A+B)D_1 + p_2(A+B)D_2) \leq g(c) = \lambda d.$$

Then, we must have $f(x_*, z_1)$ in the first case of Proposition 3 and $f(x_*, z_2)$ in the second case. Thus, indeed,

$$p_1 f(x_*, z_1) + p_2 f(x_*, z_2) = p_1 \left(\frac{AD_1 + \sqrt{(AD_1)^2 - \frac{4}{\lambda d}}}{2} \right) + p_2(A+B)D_2 = c.$$

□

3 Main theorem

Theorem 1. *Let $\alpha = 0$ and let $c < p_1AD_1 + p_2(A+B)D_2$. Suppose*

$$\frac{p_1}{p_1AD_1 + p_2(A+B)D_2 - c} \leq \frac{-1}{BD_2}.$$

Then we choose values of d: d_{def} and d_{pp} for the default and peer-prediction mechanisms respectively such that in either case

$$p_1 f(x_*, z_1) + p_2 f(x_*, z_2) = c$$

with the other parameters of the two systems being held constant. Then the term corresponding to the average penalty for the default mechanism of Proposition 2 is larger than that for the peer-prediction mechanism of Proposition 3 where x_* is the appropriate equilibrium value of effort in each case. Namely,

$$p_1 d_{def}(1 - f(x_{def,*}, z_1)) + p_2 d_{def}(1 - f(x_{def,*}, z_2))$$

$$\geq p_1 d_{pp}(1 - f(x_{pp,*}, z_1)) \cdot f(x_{pp,*}, z_1) + p_2 d_{pp}(1 - f(x_{pp,*}, z_2)) \cdot f(x_{pp,*}, z_2).$$

Proof. Notice that

$$D_2 \leq D_1 \Rightarrow p_1(A+B)D_2 + p_2(A+B)D_2 \leq p_1(A+B) + p_2(A+B)D_2 \leq c$$

where the last inequality is part of our assumptions on c. Thus,

$$\frac{p_1}{c - p_2(A+B)D_2} \leq \frac{1}{(A+B)D_2}.$$

Then the assumption:

$$\frac{p_1}{p_1 A D_1 + p_2(A+B)D_2 - c} \leq \frac{-1}{BD_2}$$

implies that the condition

$$\frac{1}{\left(AD_1 - \frac{c - p_2(A+B)D_2}{p_1}\right) \cdot \left(\frac{c - p_2(A+B)D_2}{p_1}\right)} \leq \frac{-1}{(A+B)BD_2^2}$$

also holds. Hence the assumptions of both Lemma 1 and Lemma 2 hold, so we take d_{def} and d_{pp} as given by those Lemmas respectively.

In particular, $f(x_*, z_2) = (A+B)D_2$ and $f(x_*, z_1) = \frac{c - p_2(A+B)D_2}{p1}$ for both mechanisms.

Then

$$p_1 \lambda d_{def}(1 - f(x_{def,*}, z_1)) + p_2 \lambda d_{def}(1 - f(x_{def,*}, z_2))$$

$$= \left[p_1 \left(1 - \frac{c - p_2(A+B)D_2}{p_1}\right) + p_2 \left(1 - (A+B)D_2\right)\right] \cdot \frac{p_1}{p_1 A D_1 + p_2(A+B)D_2 - c}.$$

Similarly,

$$p_1 \lambda d_{pp}(1 - f(x_{pp,*}, z_1)) \cdot f(x_{pp,*}, z_1) + p_2 \lambda d_{pp}(1 - f(x_{pp,*}, z_2)) \cdot f(x_{pp,*}, z_2)$$

$$= \left[p_1 \left(1 - \frac{c - p_2(A+B)D_2}{p_1}\right) \cdot \left(\frac{c - p_2(A+B)D_2}{p_1}\right) + p_2 \left(1 - (A+B)D_2\right) \cdot (A+B)\, D_2\right] \cdot \frac{1}{\left(AD_1 - \frac{c - p_2(A+B)D_2}{p_1}\right) \cdot \left(\frac{c - p_2(A+B)D_2}{p_1}\right)}$$

Then we see that the condition

$$p_1 d_{def}(1 - f(x_{def,*}, z_1)) + p_2 d_{def}(1 - f(x_{def,*}, z_2))$$

$$\geq p_1 d_{pp}(1 - f(x_{pp,*}, z_1)) \cdot f(x_{pp,*}, z_1) + p_2 d_{pp}(1 - f(x_{pp,*}, z_2)) \cdot f(x_{pp,*}, z_2)$$

$$\Leftrightarrow \frac{p_1(A+B)D_2}{c - p_2(A+B)D_2} \leq 1$$

$$\Leftrightarrow p_1(A+B)D_2 + p_2(A+B)D_2 < c,$$

which holds by our assumptions on c.

□

Remark 1. *Hence, we see that the peer-prediction mechanism requires smaller average penalties to achieve the same overall level of accuracy c. For participants such that $\lambda > 1$ this translates into requiring smaller fee parameters f for the expected risk-adjusted return of the participants to be positive.*

To Do

- Try to remove or weaken the

$$\frac{p_1}{p_1 A D_1 + p_2 (A + B) D_2 - c} \leq \frac{-1}{B D_2}$$

  condition. These results should intuitively hold for more general values albeit with more complicated algebra/arguments.

- Remove the $\alpha = 0$ conditions. Once one has done this one has a sense of the cost of attackers to grief so one can perform an asymptotic analysis of the griefing factors.

- Generalize to more than two difficulty levels z_1 and z_2. Ideally we should be able to accomodate difficulty distributed as a random variable, for example, given by a Weibull distribution.

References