Predicting Your Own Effort

David F. Bacon
IBM Research
dfb@watson.ibm.com

Yiling Chen
SEAS, Harvard University
yiling@eecs.harvard.edu

Ian Kash
MSR Cambridge
iankash@microsoft.com

David C. Parkes
SEAS, Harvard University
parkes@eecs.harvard.edu

Malvika Rao
SEAS, Harvard University
malvika@eecs.harvard.edu

Manu Sridharan
IBM Research
msridhar@us.ibm.com

ABSTRACT

We consider a setting in which a worker and a manager may each have information about the likely completion time of a task, and the worker also affects the completion time by choosing a level of effort. The task itself may further be composed of a set of subtasks, and the worker can also decide how many of these subtasks to split out into an explicit prediction task. In addition, the worker can learn about the likely completion time of a task as work on subtasks completes. We characterize a family of scoring rules for the worker and manager that provide three properties: information is truthfully reported; best effort is exerted by the worker in completing tasks as quickly as possible; and collusion is not possible. We also study the factors influencing when a worker will split a task into subtasks, each forming a separate prediction target.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Economics, Theory

Keywords

Information Elicitation, Proper Scoring Rule, Principal-Agent

1. INTRODUCTION

Software engineering is one of many domains with complex and modular tasks. There are often information asymmetries, both between the worker performing a task and the manager supervising and between the two of them and the rest of the organization or company. In such environments, it is important for the organization to be able to elicit accurate predictions from worker-manager teams in regard to when individual tasks are expected to complete. By eliciting accurate predictions, this enables good decision-making in regard to scheduling resources to projects (such as bug fixes or new features), and in regard to coordination of projects.

A particular challenge is that a worker with information relevant to the prediction task also controls the completion time through the amount of effort exerted on the task. In modeling this, we consider a single worker and a single manager. The worker works on a sequence of tasks and both the worker and the manager receive a score based on predictions and completion times for each task completed. We assume that the organization (or company) couples the score received by the worker or manager with incentives, be they social-psychological such as praise or visibility, or material rewards through prizes or the payment of bonuses. Based on this, we assume that the worker and the manager each seek to maximize total expected score.

The role of the worker is to share information relevant to the expected completion time of the task with the manager, in order to enable accurate predictions, and also to decide on whether to work at "best effort" or less than best effort. The role of the manager is to combine information received from the worker with her own information (if any), and make accurate predictions to the organization regarding the completion time of tasks. We tackle the issue of how to elicit truthful information and thus accurate predictions from the worker and manager, as well as how to elicit best effort from the worker.[1]

In essence, our problem is a combination of a repeated principal-agent problem and a prediction problem. In a principal-agent setting, a principal wishes to elicit a desired effort level from an agent but does not require the agent to make any predictions. On the other hand in a prediction problem, accurate predictions of the outcome of an event are sought but without considering that the distribution on outcomes might be something that can be controlled by the agent doing the prediction. In contrast we seek to establish both accuracy and the investment of best effort.

Our main technical result is a characterization of a class of scoring rules that are able to align incentives with both accurate prediction and the investment of best effort. In addition, the scoring rules inherently preclude the possibility of collusion between the worker and manager in their participation in the scoring system. For example, it is not useful for a manager and worker to agree that the worker will deliberately slow down in return for a prediction task with lower variance and thus the potential for higher total score to the worker-manager pair.

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

[1] We assume that existing incentive schemes within the organization (e.g., pay, promotion, etc.) encourage best effort work, all things being equal. For this reason, it is sufficient for our purposes that the incremental incentives provided by the scoring scheme, work with (not against) best effort. In particular, we want to preclude working at less than best effort leading to a higher expected score.

In addition, we consider the effect of a scoring system on whether or not a worker will choose to split a task into multiple prediction targets. For this purpose, we model a task as a sequence of subtasks, where a subtask is conceptualized as a unit of work with a well-defined end point, and for which the time to complete the unit of work may be informative as to the time to complete other subtasks that comprise a task. With this in mind, we study the incentives for a worker to "split-out" a subtask for the purpose of a separate prediction target.[2] The qualitative result we obtain is that there is a greater propensity to split subtasks for which the completion times are positively correlated than those for which the completion times are independent. A simulation study completes the paper, providing a quantitative analysis of the trade-off between the frequency of "splitting" prediction into subtasks, the degree to which the distribution on subtask completion time is correlated, and a parameterization of the scoring rule that affects how much payment is made per subtask target vs how much payment must be made in catch-up upon the completion of a task.

1.1 Related Work

Scoring rules have been developed to measure the performance of experts who are solicited to reveal their probability assessments regarding uncertain events. They have been used in a variety of scenarios, from weather forecasting to prediction markets [3, 5, 4, 7]. *Proper* scoring rules incentivize truthful reporting of likelihood estimates. An overview of the theory behind proper scoring rules can be found in Gneiting and Raftery [3].

Proper scoring rules typically require that the outcome of the uncertain event will be revealed and the agent whose assessment is elicited can not influence the outcome. In our setting, the prediction of effort required to complete a task and the outcome or realized effort are not independent; both are influenced by the worker. Shi *et al.* [11] consider situations where agents may be able to take actions that influence the outcome. They propose *principal-aligned* mechanisms that do not incentivize agents to take actions that reduce the utility of the principal. Their setting considers eliciting a probability distribution and the outcome space is discrete. Our setting allows for continuous effort level and we seek to elicit the expectation as well as incentivize best effort. The result of Shi *et al.* [11] can be generalized to the setting of eliciting the expectation for a random variable over a continuous outcome space using the characterization of Savage [10], which is also used to derive our characterization in Section 3. With this generalization, it is possible to derive our Theorem 7 by assigning a particular utility function to the principal and applying the result of Shi *et al.* [11]. However, this approach seems unnecessarily complicated in our setting, and we derive our results by directly considering desirable properties of the incentive mechanism.

There is a vast literature on *principal-agent* models [2, 6]. In a classical principal-agent model with hidden action, an agent chooses an action to take that is costly for him

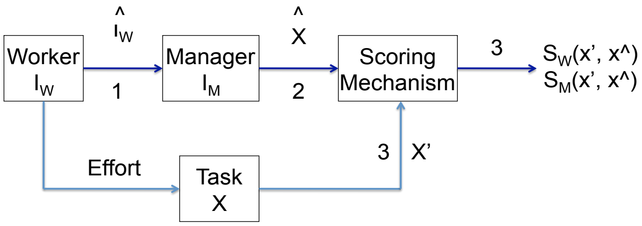

Figure 1: Timeline of the worker-manager game.

but beneficial for the principal in exchange for a promise of payment. The principal cannot directly observe the agent's action, but the stochastic correlation between actions and outcomes (that is, the probability of observing an outcome given that the agent takes an action), is common knowledge. For example, the agent's action can be a level of effort exerted with the probability of success for a project an increasing function of the level of effort. Knowing the stochastic correlation, the principal seeks to incentivize the agent to take a desirable action using contracts with payments based on the outcome.

Radner [9] considers an infinitely repeated setting for the principal-agent problem. In Radner's setting, the game is composed of sequences of review and penalty periods. By allowing the players' actions in one period to depend on the history of previous periods, the principal can observe the results of the agent's actions and *punish* the agent if the agent's performance fails some statistical test of efficiency. Radner shows that for high enough discount factors, there exist equilibria, consisting of reward-decision pairs, of the infinitely repeated game that are strictly more efficient than the short-term equilibrium. Our setting is different in that it combines the challenge of eliciting desirable actions with that of eliciting information from an agent. Our setting introduces information asymmetry about the stochastic correlation between the action and the outcome, allowing the agent to have private information about this stochastic correlation. The principal would like to elicit the information from the agent so as to obtain a better prediction, which is then used by the principal to set the reward for the agent. Because the reward of the agent now depends on the reported information, this introduces incentives to lie about the information or act in a suboptimal way. Given this tension, we aim to achieve truthful elicitation of private information as well as elicitation of the desirable action.

2. THE BASIC MODEL

We consider the incentive design problem for a company (the principal), whose goal is to truthfully elicit information from its employees as well as incentivize them to exert optimal level of effort. The basic model considers a single task with two agents, a worker and a manager, each with private information in regard to likely completion time of the task. The worker shares information with the manager, who then combines this information with his own private information and makes a prediction. The worker then exerts effort, and at some subsequent time the task completes and the worker informs the system (and the manager) of this event. We assume that only a truly completed task can be claimed as complete, but allow a worker to reduce effort below best

[2] Our viewpoint is that it is the worker, not the manager, who is privy to information in regard to subtasks. Moreover, we can imagine situations in which predictions in regard to subtasks rather than in regard to the aggregate time for a task is useful; e.g., for sharing information with other workers, for re-planning, and in order to collect data to enable the training of predictive models in order to enable better organizational efficiency going forward.

effort, including to pretend a task is not complete when it has been completed. Eventually, a score is assigned to both the worker and the manager. Later, we extend the basic model to include structure in regard to subtasks and also to consider a sequence of tasks.

Let X denote the random variable for the time the task takes to complete under the best effort by the worker. Assume that the realized value of X is nonnegative and upper bounded, i.e. $x \in (0, x_{\max})$. Neither the manager nor the worker knows the realization of X. But they each have some private information, denoted as random variables I_m and I_w respectively, on the completion time under best effort. The joint distribution $\Pi(X, I_m, I_w)$ is common knowledge to them, but not known to the company. Assuming Π is not known to the company ensures a broad range of priors are considered possible. In particular, this allows $E[X|I_m]$, $E[X|I_w]$, and $E[X|I_m, I_w]$ all take on all values in $(0, x_{\max})$. This ensures that rules we derive work for a broad range of beliefs, similar to proper scoring rules requiring truthful reporting be optimal for all probability distributions. If the company believes that only a significantly restricted set of priors is possible, there may be additional rules that our results do not characterize. Note also that these expectations are well defined because X is bounded.

The manager and the worker play a three-stage game as shown in Figure 1. In stage 1, the worker can communicate with the manager and share information. In stage 2, the manager makes a prediction $\hat{x}$ about the completion time of the task under the worker's best effort. In stage 3, the worker exerts some effort and completes the task in time x'. While the worker cannot exert more than his best effort and complete the task in time less than x, he can work at a more slack pace and take time $x' > x$ to complete the task. However, we require that $x' \leq x_{\max}$ because otherwise it will be clear to both the manager and company that he is not working efficiently.

We assume that both the manager and the worker are risk neutral. We further assume that the worker, all things being equal, is indifferent between working at best effort or "slacking." In other words, if the worker can get a higher expected score through a best-effort strategy rather than slowing down, then this is the approach the worker will take. Our results also hold when there is an existing, strict incentive for best effort over slacking, for example because of existing incentives in the company.

We consider incentive mechanisms (we refer to them as *scoring systems*) that reward the manager and the worker based on the manager's prediction of the completion time and the worker's actual completion time. At the end of stage 3, a manager is rewarded according to the score $S_m(x', \hat{x})$ and the worker according to the score $S_w(x', \hat{x})$. We require S_m and S_w to be differentiable with respect to x' and $\hat{x}$. The goal of a scoring system is to incentivize the report of an accurate prediction at best effort and the exertion of the best effort.

2.1 Desirable Properties of Scoring Systems

Our model is a simple two-player three-stage game. We hence consider the perfect Bayesian equilibrium of the game and desire good behavior of the manager and the worker at the equilibrium. The following are four properties we would like a scoring system to achieve at the equilibrium:

1. **Information sharing**. For all Π, the worker shares his private information I_w honestly with the manager in stage 1.
2. **Report the mean**. For all Π, when estimating the time required to complete a task under best effort of the worker, the manager's optimal report in stage 2 is $\hat{x} = E[X|I]$ where I is all information available to the manager at the time, given equilibrium beliefs.
3. **Best effort**. For all $\hat{x}$, it is optimal for the worker to exert his best effort and choose $x' = x$ for all realizations x in stage 3.
4. **Collusion-proofness**. For all Π, the total expected score of the manager and the worker is maximized by reporting $\hat{x} = E[X|I_w, I_m]$ and exerting best effort such that $x' = x$ for all realizations x.

If the above four properties are satisfied, we will have a perfect Bayesian equilibrium where the worker shares all his information with the manager, the manager truthfully reports his expectation of the completion time under best effort given both pieces of information, and the worker completes the task as quickly as possible. Moreover, this equilibrium is collusion-free, such that no joint deviation can lead to an increase in the total expected score.

3. CHARACTERIZATION OF SCORING SYSTEMS

We proceed to characterize scoring systems that satisfy our desirable properties. The main technical challenge is to simultaneously address the need for accurate prediction and retain incentives for the worker to adopt best effort.

First, we consider the best effort property. It's easy to see that if choosing $x' = x$ is optimal for the worker given any x and prediction $\hat{x}$, the worker's score $S_w(x', \hat{x})$ must be a decreasing function of x'.

OBSERVATION 1. *A scoring system satisfies best effort if and only if* $\frac{\partial S_w(x', \hat{x})}{\partial x'} \leq 0$.

For example, a simple scoring rule $S_w(x', \hat{x}) = 2\hat{x} - x'$ can incentivize the worker to exert his best effort.

Given the best effort property, we know that x' is set to x at the equilibrium. The report the mean property requires a scoring system to incentivize the manager to honestly report her expected completion time given all available information. This is exactly the problem addressed by proper scoring rules for eliciting the mean of a random variable. Proper scoring rules for eliciting the mean of a random variable satisfy the property that reporting the mean maximizes expected score. Hence, we have an immediate solution based on the definition of proper scoring rules.

OBSERVATION 2. *If the best effort property is satisfied, the scoring system satisfies the report the mean property if and only if* $E(X|I) \in \operatorname{argmax}_{\hat{x}} E(S_m(X, \hat{x})|I)$.

We can use any proper scoring rule as the manager scoring rule, in conjunction with a worker scoring rule that incentivizes best effort, to achieve the report the mean property. For example, $S_m(x', \hat{x}) = b - (\hat{x} - x')^2$ for an arbitrary parameter b uses a quadratic scoring rule.

While it is easy to achieve both best effort and report the mean properties at an equilibrium, satisfying information sharing and collusion-proofness is less straightforward.

Consider the pair of the worker and manager scoring rules mentioned above, $S_w(x',\hat{x}) = 2\hat{x} - x'$ and $S_m(x',\hat{x}) = b - (\hat{x} - x')^2$. The worker may not want to share his information with the manager if his information will lead to a lower prediction $\hat{x}$ by the manager. In addition, the total score can be increased if the worker and the manager collude. To see this, note that the manager can report a larger prediction and the worker can work slowly to perfectly match the manager's prediction, which increases the worker's score while maximizing the manager's score. Below, we characterize the conditions for achieving all four desired properties simultaneously.

3.1 A Family of Scoring Rules

We first consider how to satisfy the information sharing property. This will require that the worker is also rewarded for a more accurate prediction.

Lemma 3. *If the best effort and report the mean properties are satisfied, the information sharing property is satisfied if and only if $E(X|I) \in \operatorname{argmax}_{\hat{x}} E(S_w(X,\hat{x})|I)$.*

Proof. The worker can influence the prediction $\hat{x}$. In an extreme case, when all relevant information is possessed by the worker, the prediction is effectively made by the worker. In order for the worker to predict the mean, the worker scoring rule needs to be a proper scoring rule for the random variable X. Because $E(X|I)$ maximizes a worker's score given any information set I, for any I_m and I_w, $E(X|I_w, I_m)$ maximizes the worker's expected score $E(S_w(X,\hat{x})|I_w, I_m)$. Hence, the worker is better off sharing the information with the manager to have the manager report $E(X|I_w, I_m)$. □

Next, we consider achieving collusion-proofness. Let $S_T(x',\hat{x})$ denote the sum of the worker and manager scores. If the manager and the worker collude to report a prediction $\hat{x}$ and complete the task in time x', collusion-proofness requires that the manager-worker pair is incentivized to report the mean and exert best effort. These are analogous to achieving information sharing and best effort when the worker has all information and the manager has no information. Let $S_T(x',\hat{x}) = S_w(x',\hat{x}) + S_m(x',\hat{x})$ be the total scoring rule. The following result follows immediately.

Lemma 4. *Collusion-proofness is satisfied if and only if $\frac{\partial S_T(x',\hat{x})}{\partial x'} \leq 0$ and $E(X|I) \in \operatorname{argmax}_{\hat{x}} E(S_T(X,\hat{x})|I)$.*

This means that if a scoring system satisfies best effort, report the mean, and information sharing we essentially get collusion-proofness for free with the mild additional condition that the total scoring rule also satisfies best effort (a sufficient condition for which is that the manager's scoring rule satisfies best effort). Combining the results characterizes scoring systems that satisfy all four desirable properties.

Lemma 5. *A manager-worker scoring system satisfies information sharing, report the mean, best effort, and collusion-proofness at a perfect Bayesian equilibrium if and only if the following conditions are satisfied:*

- $\frac{\partial S_w(x',\hat{x})}{\partial x'} \leq 0.$
- $\frac{\partial S_T(x',\hat{x})}{\partial x'} \leq 0.$
- $E(X|I) \in \operatorname{argmax}_{\hat{x}} E(S_m(X,\hat{x})|I).$
- $E(X|I) \in \operatorname{argmax}_{\hat{x}} E(S_w(X,\hat{x})|I).$

for all information sets I.

Intuitively, Lemma 5 requires that the worker score and the manager score are all given by a proper scoring rule for eliciting the mean (it is immediate that the total score must also be given by a proper scoring rule), in addition to the worker and total scores being a decreasing function of the actual completion time. For example, $S_w(x',\hat{x}) = S_m(x',\hat{x}) = f(x') + 2cx'\hat{x} - c\hat{x}^2$, where $f'(x') + 2c\hat{x} < 0$ and $c > 0$ is a family of scoring systems that satisfy all four desirable properties. A theorem due to Savage [10] characterizes all (differentiable) proper scoring rules for eliciting the mean.

Theorem 6 (Savage [10]). *For S differentiable in $\hat{x}$, $E(X|I) \in \operatorname{argmax}_{\hat{x}} E(S(X,\hat{x})|I)$ if and only if $S(x',\hat{x}) = f(x') + G(\hat{x}) + (x' - \hat{x})G'(\hat{x})$ where $E[f(X)|I]$ is finite for all Π and G is a differentiable convex function.*

Note that a sufficient condition for $E[f(X)|I]$ to be finite for all Π is that f is bounded on $(0, x_{\max})$. Combining Theorem 6 with Lemma 5 yields a more precise characterization.

Theorem 7. *A manager-worker scoring system satisfies information sharing, report the mean, best effort, and collusion-proofness at a perfect Bayesian equilibrium if and only if the following conditions are satisfied:*

- *$S_w(x',\hat{x}) = f_w(x') + G_w(\hat{x}) + (x' - \hat{x})G'_w(\hat{x})$ where f_w is a differentiable function such that $E[f_w(X)|I_w]$ is finite for all Π and G_w is a differentiable convex function.*
- *$S_m(x',\hat{x}) = f_m(x') + G_m(\hat{x}) + (x' - \hat{x})G'_m(\hat{x})$ where f_m is a differentiable function such that $E[f_m(X)|I_m, I_w]$ is finite for all Π and G_m is a differentiable convex function.*
- *$f'_w(x') + G'_w(\hat{x}) \leq 0$ for all x',$\hat{x} \in (0, x_{\max})$.*
- *$f'_w(x') + f'_m(x') + G'_w(\hat{x}) + G'_m(\hat{x}) \leq 0$ for all x',$\hat{x} \in (0, x_{\max})$.*

Finally, note that this means we can derive a scoring system from a differentiable convex pair of Gs whose derivatives we can upper bound by taking $f'_w(x') = -|\sup_{\hat{x}} G'_w(\hat{x})|$ and similarly for f_m.

4. TASK DECOMPOSITION

Continuing, we now consider that a task has substructure, with a task represented as a series of subtasks. Based on this, we allow a worker-manager team to elect to split-off individual subtasks (or contiguous subtasks) to become identified prediction tasks in their own right; i.e., essentially partitioning the task into a distinct set of pieces, each of which has an associated prediction problem.

In increasing the realism of the model, we also situate the prediction task for a single task in the context of a repeated version of the problem, in which a worker has a sequence of tasks. In this context, the following property is useful:

5. **Always non-negative**. The score of the worker and the manager is always non-negative for all realizations of x and all reports $\hat{x}$.

If the score is always non-negative, then our best effort property immediately guarantees that best effort is also optimal for a worker facing a sequence of tasks, in that this will maximize both the total score for sequence of tasks and the score per unit time.[3]

This noted, we can focus back on a single task and introduce formalism to make precise what is intended by a subtask. Let $X = X_1 + \ldots + X_k$ denote a task X composed of k subtasks $X_1, \ldots, X_k$. The worker decides which sets of subtasks are to become targets of the scoring system. For example, the worker might prefer to make a single prediction, thereby being scored just once after completing the task in its entirety. Another option is that the worker may prefer to make k predictions (hence receiving k scores), one for each subtask. Alternately, the worker select subtask X_1 as a target, then subtask X_2, and then subtasks $X_3, \ldots, X_k$ aggregated into one chunk of work for the purpose of prediction. We assume that the degree to which the prediction problem associated with a task may be split-out into subtasks is knowledge that is private to the worker and *a priori* not known to the manager.

We allow the worker to make online decisions about which subtasks to split-out as separate prediction targets. That is, if the worker initially decides to get scored for X_1, after this is done he can then choose whether to next get scored for X_2 or instead to combine X_2 with some number of subsequent subtasks (we assume subtasks must be completed in order). As we are focusing on decisions made by the worker, we will only discuss S_w. The report the mean and collusion proofness properties can be retained through an appropriate choice of S_m. To be able to make concrete statements, we focus on the special case $S_w(x, \hat{x}) = f(x') + 2cx'\hat{x} - c\hat{x}^2$.

4.1 Independent Subtasks

For a simple model, consider a worker with two subtasks, denoted by random variables X_1 and X_2, and each with discrete support $\{a, b\}$, with $0 < a < b \leq 1$ and $x_{\max} = b$.

For this setting with two subtasks, the choice of the worker in regard to prediction targets is as follows:

- Adopt the complete task as a prediction target, share information in regard to $X = X_1 + X_2$ (with the manager making a prediction), work on them both, and then receive a score.

- Split-out X_1 as the first prediction target, share information with the manager (with the manager making a prediction), work on X_1 and receive a score, then share information in regard to X_2, work and receive a score.

Lemma 8. *Let $S_w(x, \hat{x}) = f(x') + 2cx'\hat{x} - c\hat{x}^2$ satisfy best effort and always non-negative. Then for a task with two subtasks, it is always optimal for the worker to split independent subtasks into separate prediction targets.*

Proof. For any distribution of effort X the worker's expected score from truthful reporting (which is optimal) is

$$E[S_w(X, E[X])] = E[f(X)] + cE[X]^2.$$

To deal with $E[f(X)]$, we make use of two bounds regarding $f(x)$. First, we know that $f'(x') < -2c\hat{x}$ for all $\hat{x}$, so in particular this is true for $\hat{x} = x_{\max}$. By always non-negative, $f(x_{\max}) \geq 0$. Thus, $f(x) \geq (x_{\max} - x)2cx_{\max}$. Second, for $a < b$, $f(a) - f(b) \geq (b - a)2cx_{\max}$. We now show that $E[S_w(X_1, E[X_1])] + E[S_w(X_2, E[X_2])] > E[S_w(X_1 + X_2, E[X_1 + X_2])]$. Note that we use the unconditional expectation over X_2 here because X_1 and X_2 are independent.

$$\begin{aligned}
&E[S_w(X_1, E[X_1])] + E[S_w(X_2, E[X_2])] \\
&- E[S_w(X_1 + X_2, E[X_1 + X_2])] \\
&= E[f(X_1)] + cE[X_1]^2 + E[f(X_2)] + cE[X_2]^2 \\
&- E[f(X_1 + X_2)] - cE[X_1 + X_2]^2 \\
&= E[f(X_1) + f(X_2) - f(X_1 + X_2)] - 2cE[X_1]E[X_2] \\
&\geq E[(x_{\max} - X_1)2cx_{\max} + ((X_1 + X_2) - X_2)2cx_{\max}] \\
&- 2cE[X_1]E[X_2] = 2c(x_{\max}^2 - E[X_1]E[X_2]) > 0.
\end{aligned}$$

□

We take this as a negative observation, because there is no learning effect when splitting out independent subtasks— it is not the case that additional accuracy can be achieved through separate predictions in the absence of correlations.

On the other hand, if we are willing to accept a scoring rule that may be negative, it is easy to obtain a different result. For example, take $f(x') = -kx'$ ($k > 2x_{\max}$) and $c = 1$. Some algebra shows that not splitting results in an increase in utility of $2E[X_1]E[X_2] > 0$, and so independent subtasks are not split out as separate prediction targets.

For this reason, the following is a very helpful observation. If the distinction between the completion of a task and the completion of a subtask is observable by the company, then the scoring system can provide a large enough *bonus score* $B > 0$ upon the completion of a task (but not a subtask), in order to remove the broader implications of a stream of negative scores. We adopt this approach going forward, allowing for scoring rules that may be negative but correcting for this with a large enough catch-up bonus B on the completion of a complete task.[4]

Parameter B can be calculated as the negation of the lowest possible score (the most negative score) that a worker who exerts best effort can possibly get for completing the task. For a given chunk of work (a set of subtasks chosen as a prediction target), the lowest score is achieved when the time to complete it under best effort is maximized while the prediction of the completion time is minimized.

4.2 Correlated Subtasks

To gain a qualitative understanding of the effect of our scoring rules on the propensity to split-out subtasks as separate targets, we adopt a simple model of correlation. The joint distribution on (X_1, X_2) is parameterized with $q \in (1/2, 1]$ and $r \in [0, 1]$. The distribution on time to complete task 1 under best effort is a with probability q and b with probability $1 - q$. With probability r, the time to complete task 2 is the same as for task 1 (i.e. $X_2 = X_1$). Otherwise, with probability $1 - r$ the time to complete task 2 is independently sampled according to probability q.

We use the scoring rule $S_w(x, \hat{x}) = f(x') + 2cx'\hat{x} - c\hat{x}^2$ with $f(x') = C - kx'$ and $c = 1$, where $k > 2x_{\max}$ and C is

[3] In contrast, suppose the score assigned for the completion of a task is negative. In this case, a worker may prefer to spend 10 hours and earn a score of -2 than to spend 1 hour and earn a score of -1, because in those additional 9 hours the worker would be completing additional tasks for more negative scores.

[4] This bonus is invariant to any aspect of the prediction or effort and does not change the rest of the analysis.

a constant. We show that, for appropriate choice of C, the incentive to split-out subtasks increases as r increases, and thus as there is more positive correlation between the time to complete the subtasks under best effort.

In particular, the choice of C sets a threshold for r. If r is below this threshold then the subtasks are independent enough that the worker does not want to split them. If r is above this threshold then the substasks are correlated enough that splitting them to learn is worthwhile. Increasing C decreases this threshold, but increases the cost to the scoring rule. Thus the choice of C allows a trade-off between encouraging the accurate sharing of predictions on subtasks and cost. However, past a certain point, the worker will want to split-out all subtasks regardless, and increasing C will simply increase the cost.

Lemma 9. *Consider a task with two sub-tasks. Let S_w be as above with $C < 2E[X_1]E[X_2]$. Let $r^* = \sqrt{\frac{2E[X_1]E[X_2]-C}{q(1-q)(a-b)^2}}$. If $r \geq r*$ then it is optimal for the worker to split-out subtasks. If $r \leq r^*$ then it is optimal for the worker to not do so.*

Proof. Unlike in Lemma 8, X_1 and X_2 are no longer independent. In particular, this means that the expected score for task two if they are split is no longer simply $E[S_w(X_2, E[X_2])]$. Instead, the worker learns something after completing the first task so, a priori, the expected score is $E_{X_1}[E[S_w(X_2, E[X_2|X_1 = x])|X_1 = x]]$. Hence we can write the expected gain from splitting as follows:

$$
\begin{aligned}
&E[S_w(X_1, E[X_1])] + E_{X_1}[E[S_w(X_2, E[X_2|X_1 = x])|X_1 = x]]\\
&- E[S_w(X_1 + X_2, E[X_1 + X_2])]\\
&= E[C - kX_1 + E[X_1]^2]\\
&+ E_{X_1}[E[C - kX_2 + E_{X_1}[(E[X_2|X_1 = x])^2]|X_1 = x]]\\
&- E[C - k(X_1 + X_2) + E[X_1 + X_2]^2]\\
&= C + E[X_1]^2 + E_{X_1}[(E[X_2|X_1 = x])^2] - E[X_1 + X_2]^2\\
&= C + E[X_1]^2 + E[X_2]^2 - E[X_1 + X_2]^2\\
&+ E_{X_1}[(E[X_2|X_1 = x])^2] - E[X_2]^2\\
&= C - 2E[X_1]E[X_2] + E_{X_1}[(E[X_2|X_1 = x])^2] - E[X_2]^2.
\end{aligned}
$$

For the particularly simple distribution we have chosen, we can expand the last two terms as

$$
\begin{aligned}
&E_{X_1}[(E[X_2|X_1 = x])^2] - E[X_2]^2\\
&= q(ra + (1-r)E[X_2])^2 + (1-q)(rb + (1-r)E[X_2])^2 - E[X_2]^2\\
&= (1-r)^2E[X_2]^2 + qr^2a^2 + (1-q)r^2b^2 + 2qar(1-r)E[X_2]\\
&+ 2(1-q)br(1-r)E[X_2] - E[X_2]^2\\
&= (1-r)^2E[X_2]^2 + qr^2a^2 + (1-q)r^2b^2 + 2r(1-r)E[X_2]^2\\
&- E[X_2]^2\\
&= qr^2a^2 + (1-q)r^2b^2 - r^2E[X_2]^2\\
&= r^2(qa^2 + (1-q)b^2 - (qa + (1-q)b)^2)\\
&= r^2(qa^2 + (1-q)b^2 - q^2a^2 - (1-q)^2b^2 - 2q(1-q)ab)\\
&= r^2((1-q)qa^2 + q(1-q)b^2 - 2q(1-q)ab)\\
&= r^2q(1-q)(a-b)^2
\end{aligned}
$$

Thus, splitting is optimal if and only if $C - 2E[X_1]E[X_2] + r^2q(1-q)(a-b)^2 \geq 0$. Solving for r yields the desired inequalities. □

In a situation where it is important that the cumulative score for each task is non-negative, then a mitigating aspect of this trade-off is that for smaller values of C the scoring system must assign a larger bonus B upon task completion to correct for the possibility of an accumulation of negative scores on subtasks.

Remark: One might also wonder whether it is possible to modify our scoring rules to allow a worker-manager team to "push" new predictions in regard to a particular prediction target over time, and without leading to new strategic considerations. For example, suppose the worker elects not to split-off any subtasks and have as the target the entire task. But now as work is completed on each subtask, perhaps the worker has updated information in regard to when the task will likely be completed. Perhaps surprisingly, the effect of allowing this turns out to be quite subtle.

For example, associating the score with the average of the score from m predictions fails, because the worker-manager team could maximize its realized score by simply pushing a lot of predictions just before completing a task when there is high confidence about how long the task will take. Insisting that predictions are made at fixed intervals of time could lead to a preference for slacking in order to be able to make an additional prediction. Adopting a time-averaged score, integrated over the different predictions made over time in regard to a prediction target, could lead to a preference to work more slowly on subtasks about which the prediction is higher quality. We leave a full reconciliation of this problem to future work.

5. SIMULATIONS

Our simulation study is designed to validate the three qualitative observations in our theoretical analysis: (a) for subtasks with more correlation the worker will tend to split out more subtasks as targets, (b) for a higher value of C the worker will tend to split out more subtasks into targets, and (c) for a higher value of C the average score received by the worker will tend to increase.

For this purpose, we consider a task X with 3 subtasks X_1, X_2, and X_3. With probability q the task is *low* difficulty, and with probability $1-q$ the task is *high* difficulty. Given that the task is low difficulty, then a subtask takes time $a = 0.5$ under best effort with probability $p \in [0.5, 1]$, and $b = 1$ under best effort otherwise. For a high difficulty task, a subtask takes time $b = 1$ with probability p, and $a = 0.5$ otherwise (both under best effort.) In this way, p controls the correlation between effort on subtasks. High p yields high correlation.

We simulate each possible policy a worker might adopt in deciding which subtasks to split-off into separate prediction targets. Altogether, there are six possible policies:

1. Policy 1: Work on each subtask separately. First target is subtask X_1, then X_2, followed by X_3.

2. Policy 2: First target is X_1. If completion time of X_1 is observed to be a then the second target is X_2, followed by X_3. If completion time of X_1 is b then the second target is $X_2 + X_3$ as a chunk.

3. Policy 3: First target is X_1. If completion time of X_1 is observed to be b then the second target is X_2, followed by X_3. If completion time of X_1 is a then the second target is $X_2 + X_3$ as a chunk.

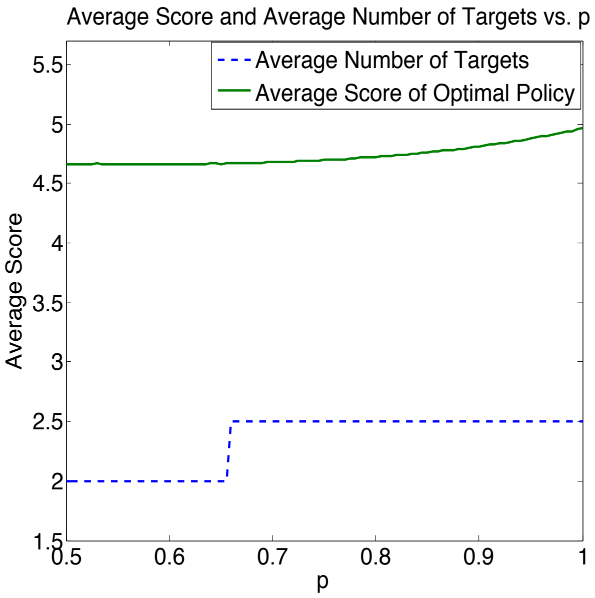

Figure 2: Average score and average number of prediction targets under the best policy, varying $p \in [0.5, 1]$ for $C = -1.9$ and $q = 0.5$.

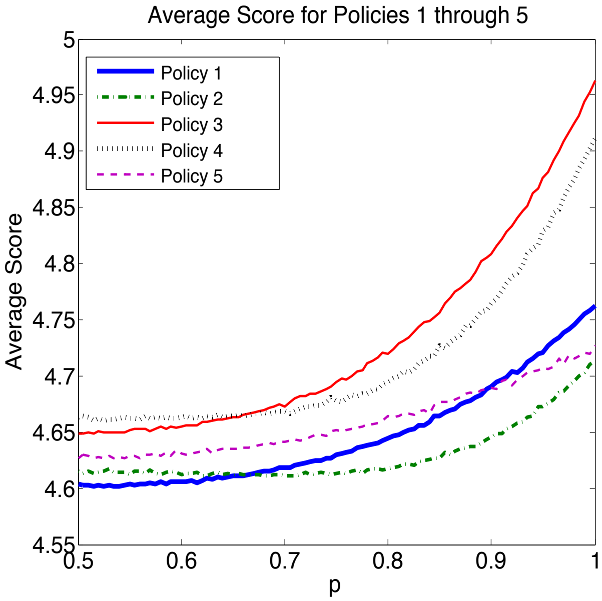

Figure 3: Average score for policies 1 through 5, varying $p \in [0.5, 1]$ for $C = -1.9$ and $q = 0.5$.

4. Policy 4: First target is X_1. The second target is $X_2 + X_3$ as a chunk.

5. Policy 5: First target is $X_1 + X_2$ as a chunk. The second target is X_3.

6. Policy 6: The first and only target is the entire task $X = X_1 + X_2 + X_3$ as a single chunk.

For concreteness, the scoring rule that we adopt is

$$S_w(x, \hat{x}) = C - 2x' x_{\max} + 2x' \hat{x} - \hat{x}^2$$

In considering the score, we also allocate a bonus B upon completion of the entire task, set to the minimal value such that the score is guaranteed to be positive for all contingencies. To determine this value, we first compute all the possible different scores that could be obtained for each policy, selecting the lowest score as that policy's worst score. The (negated) lowest score amongst the 6 worst scores of the 6 policies provides the bonus.

Given this setup, we compare the average score and the average number of prediction targets as the amount of positive correlation (reflected by p) and the parameter in the scoring rule C varies. For each policy, and for different values of C, p and q, we run at least 10,000 trials and determine the average score. The policy that we assume the worker adopts for a triple (C, p, q) is that which maximizes the average score.

Figure 2 is obtained by varying $p \in [0.5, 1]$ for $C = -1.9$ and $q = 0.5$, and shows for each value of p the average score and the average number of targets for the optimal policy *for that value of* p. As p increases there is greater correlation which results in more splitting and a higher score. Figure 3 corroborates this by showing that as p approaches a value of 0.7, the optimal policy changes from Policy 4 to Policy 3. Since Policy 3 varies between 2 and 3 splits, we get an average number of targets equal to 2.5. We have omitted Policy 6 in Figure 3; its average score remained in the range $[2.8, 2.9]$ throughout.

Figure 4 is obtained by varying $C \in [-5, 1]$ for $p = 0.8$ and $q = 0.5$, and shows for each value of C the average score and the average number of targets for the optimal policy *for that value of* C. It shows that as C increases there is more splitting. while the average score also increases. Figure 5 shows the average score of the different policies. The optimal policy is initially Policy 6 (no splitting), and hence there is only 1 prediction target. With increasing C the optimal policy changes to those with greater splitting, finally ending up at Policy 1 (full splitting). We have omitted policies 2 and 5 in Figure 5, as their scores were very close to the scores of policies 3 and 4 respectively (policies 2 and 5 scored slightly less than policies 3 and 4 respectively for all values of C). For $C < -3.75$, the value of B was determined by policy 1, which is why the curve for policy 1 is initially flat and the others are decreasing. For larger values of C, B is determined by policy 6 so it is flat while the others increase.

The basic trends we see in these plots are consistent with the theory, which allows for a tradeoff between the degree to which tasks are split and the cost to the mechanism.

6. CONCLUSIONS

We have introduced the problem of incentivizing a worker-manager team to commit best effort to a task and make accurate predictions in regard to completion time. In studying this question, we have characterized a family of scoring rules with natural properties, and considered the effect of the rules on decisions in regard to which subtasks to split-out into explicit prediction targets.

The problem was motivated by an extant outcomes-based incentive system currently applied to IBM's internal IT initiatives. In this system, software professionals (developers, software designers, testers, etc.) execute tasks assigned by their project managers to produce project deliverables. Each

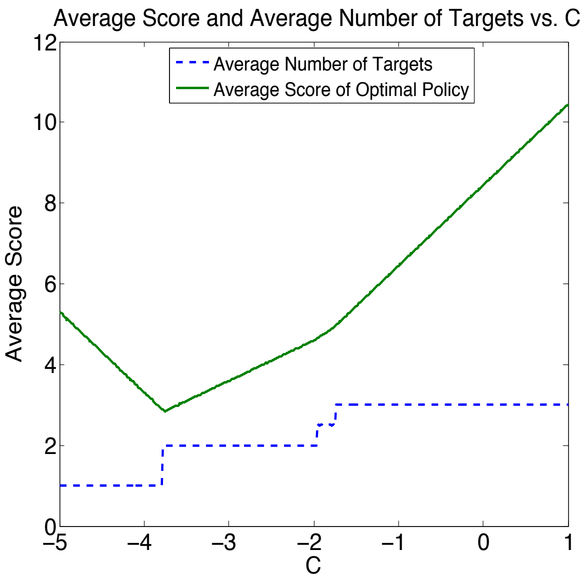

Figure 4: Average score and average number of prediction targets under the best policy, varying $C \in [-5, 1]$ for $p = 0.8$ and $q = 0.5$.

task is associated with a "Blue Sheet" that records the manager's prediction of required effort for the task, along with its actual completion time. Blue Sheet data are used to compute scores for both 'workers' and 'managers,' and top scorers are recognized for their achievement.

The Blue Sheet system has been in place since 2009 and has provided some useful initial insights on process differences across internal groups. However, the current Blue Sheet scoring system does not satisfy any of the four properties outlined in Section 2.1. It is difficult to derive any strong conclusions about the impact of these missing properties from existing Blue Sheet data (much of the information is self-reported), but the data suggests some evidence of collusion between 'workers' and 'managers.'

We are aiming to pilot a new scoring system based on the current work, comparing to the existing system, both by comparing scores and outcomes and by surveying the participants regarding which system they prefer. It will be interesting to consider, as a next step, additional factors that might be important in a practical deployment. These factors include the impact of a scoring system on the kinds of tasks that worker-manager teams choose to take on, for instance in regard to their inherent predictiveness.

The current Blue Sheet system includes some additional aspects that are outside of our model. These include a self-assessment of the deliverable quality against specified standards, and also an assessment of the extent that re-use of pre-existing assets was leveraged to complete the deliverable. From this perspective, we are interested to understand the impact of a scoring system on how to decompose work into subtasks in the first place, that is on the modularization of tasks. A key goal of the Blue Sheet system is to incentivize the creation and application of reusable software components, thereby making the development process more efficient. Devising incentive schemes that directly encourage creating reusable components, e.g., by rewarding the component author when others reuse the component, remains as future work.

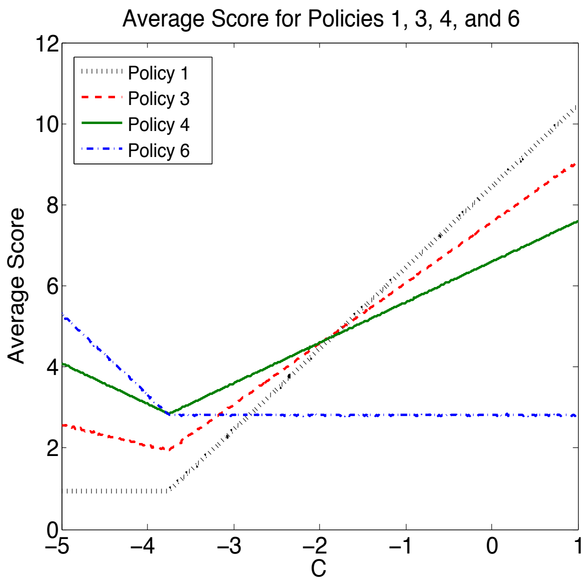

Figure 5: Average score for policies 1, 3, 4, and 6, varying $C \in [-5, 1]$ for $p = 0.8$ and $q = 0.5$.

7. ACKNOWLEDGMENTS

This material is based upon work supported by NSF Grant No. CCF-0953516. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors alone.

8. REFERENCES

[1] D. F. Bacon, E. Bokelberg, Y. Chen, I. A. Kash, D. C. Parkes, M. Rao, and M. Sridharan. Software Economies. In *Proc. FSE/SDP Workshop on the Future of Software Engineering Research*, 2010.

[2] D. Fudenberg, and J. Tirole. *Game Theory*. MIT Press, 1991.

[3] T. Gneiting, and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 102(477):359–378, March 2007.

[4] R. D. Hanson. Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets*, 1(1):1–15, 2007.

[5] N. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proc. 9th ACM Conf. on Electronic commerce* (EC '08), pp. 129–138, 2008.

[6] A. MasColel, M. D. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[7] A. H. Murphy and R. L. Winkler. Probability forecasting in meteorology. *J. Am. Stat. Assoc.*, 79(387):489–500, 1984.

[8] R. Radner. Monitoring Cooperative Agreements in a Repeated Principal-Agent Relationship. *Econometrica*, Vol. 49, No. 5, pp. 1127– 1148, September 1981.

[9] R. Radner. Repeated Principal-Agent Games with Discounting. *Econometrica*, Vol. 53, No. 5, pp. 1173–1198, September 1985.

[10] L. J. Savage. Elicitation of personal probabilities and expectations. *J. Am. Stat. Assoc.*, 66(336):783–801, 1971.

[11] P. Shi, V. Conitzer, and M. Guo. Prediction mechanisms that do not incentivize undesirable actions. In *Proc. 5th International Workshop on Internet and Network Economics* (WINE '09), pp. 89–100, 2009. Springer-Verlag.