Incentivizing Effort and Ensuring Impartiality on Amazon Mechanical Turk

Anson Kahng
Computer Science Department
Carnegie Mellon University
akahng@cs.cmu.edu

Yasmine Kotturi
Human-Computer Interaction Institute
Carnegie Mellon University
ykotturi@cs.cmu.edu

1 Abstract

Expert crowd work faces a fundamental challenge. While platforms like Upwork allow employers to access specialized labor, they require employers to have domain knowledge to assess applicants manually. This slows hiring and limits the kinds of employers who can use expert crowdsourcing. One current approach aims to leverage the domain expertise of the applicants through impartial peer ranking and aggregation, an approach that exploits the fact that the voters are also the set of alternatives and ensures the property that no voter can influence her position in the final ranking. Through leveraging domain expertise of voters (i.e. crowd-expert job applicants) and ensuring impartiality, this approach is a step in a promising direction for hiring in online labor markets. Importantly, such impartial mechanisms do not incentivize strategic behavior (i.e. are strategyproof from the perspective of self-interested voters); however, they do not incentive effort. We therefore extend prior work to incentive effort whilst maintaining impartiality. In particular, we show that for a broad class of bonus-based schemes, it is possible to ensure both impartiality and incentivize effort when bonuses are small, and it is possible to ensure a new property we refer to as *effort-impartiality* when bonuses are large. In a between-subjects experiment on Amazon Mechanical Turk (N=90), participants 1. receive no additional bonus for effort, 2. receive a small bonus, or 3. a large bonus, relative to base-rate payment. We hypothesize that the larger effort bonus leads to increased effort, and by extension accuracy (as participants are task-domain experts). Our findings suggest a two-fold consequence of introducing effort payments: 1. both perceived effort and time on task increase when effort payment is highest and surprisingly, 2. increased effort does not lead to increased accuracy, as compared to expert ground truth.

2 Introduction

Online outsourcing enables many individuals and businesses to hire experts for particular tasks, and for short periods of time (11). This flexibility and access to hard-to-find expertise has profound benefits, and is reflected in its rising popularity: the number of hours worked increased 55% from 2011 to 2012, and workers earned over $360 million (2). This growth has kept up in recent years: for employers, expert online outsourcing provides broader access to specialized skills and 24hour productivity; for workers, this has created new opportunities to access and compete in global job markets, from anywhere at any time, as long as they have computer and Internet access (11).

Online labor platforms today require employers to evaluate applications from expert crowdworkers to make hiring decisions. However, assessing workers' applications accurately requires domain expertise. This lack of domain expertise prevents employers from hiring workers in areas where they lack expertise (Consider a hairdresser hiring a web designer to make the business website.) Furthermore, because employers are unable to distinguish workers, online expert crowdsourcing markets become "markets for lemons" over time (3): requesters offer lower wages to offset their risk of low-quality results, and workers respond to lower payment with lower quality work (17).

Final project report for course CMU 10709 "Fundamentals of Learning from the Crowd", Fall 2017. Instructor: Nihar B. Shah, TA: Ritesh Noothigattu. The formatting style file is borrowed from the NIPS conference.

These lower wages discourage qualified workers from participating, in turn driving away potential employers. Existing approaches that aim to inform employers of worker quality (such as reputation systems) have fundamental flaws that prevent them from reflecting worker quality accurately (8). Even if employers do possess some domain expertise that they can leverage to hire workers, they still have to deal with the significant search friction. On online expert outsourcing platforms like Upwork, it takes employers three days to screen, interview, and hire candidates (upw).

One recent approach to this problem leverages the concept of *impartiality*, a property that guarantees that each worker has no effect on her final ranking, in order to develop strategyproof aggregation methods with accuracy guarantees. In particular, recent work by (9) extends the theory of impartial rank aggregation mechanisms, which take as input comparisons in which the set of people being compared and the set of people doing the comparisons are the same, and which output a social ordering over the set of all participants. The mechanisms satisfy *impartiality*, which means that each player should not be able to influence their own rank in the final output. However, although these impartial algorithms satisfy impartiality and provide various "closeness" guarantees of the final result, there is one issue that remains to be addressed. In its current form, there is no incentive for players to put in the effort to provide accurate comparisons of their peers. In fact, communication of the presence of an impartial algorithm leads to voter perception that effort invested has no benefit whatsoever, and thus there is less optimal performance compared to no mention of impartial algorithm (10). Therefore, we focus on this incongruency of prior work:

Problem Statement: Impartiality guarantees that each player will not be able to benefit from reporting incorrect results, but it does not incentivize the player to put in any effort to report correct results. We aim to address this misalignment of incentives by designing a bonus-based payment mechanism that is both impartial and effort-inducing.

2.1 Related Work

Our approach involves an approach similar to that of (16) in which we can insert known ground truth queries and hopefully use a rule that relies on a thresholded discriminator that takes advantage of concentration bounds similar to the one on mini-homework 5 to incentivize effort. However, for schemes that do not involve the insertion of synthetic ground truth (or the elicitation of a small sample of ground truth via, e.g., TA or other expert evaluations), it is much harder to ascertain what constitutes "ground truth". Recent work by (5) addresses this by extending the Bradley-Terry model and formulating this as a regularized maximum likelihood problem. One approach is using naive output agreement between players in which players are penalized for, in a sense, "going against the grain". Another related approach involves using the Bayesian truth serum by asking them not only for pairwise comparisons, but also for their estimates of how many people will agree with each of their comparisons (15) (19), and perhaps it could be interesting to examine the theoretical guarantees (or impossibility results) we obtain from these forms of approximating ground truth from unknown comparisons.

However, although we initially set out to address two settings: one in which there exists limited access to ground truth and one in which there does not exist any ground truth, further reading led us to focus on the limited ground truth setting. In particular, recent work by Gao et al. (6) came to the surprising conclusion that, when incentivizing effort on the part of participants, a very naive reward mechanism based on limited access to ground truth led to much better incentive guarantees than state-of-the-art peer prediction methods. In particular, the authors concluded that when there exist multiple signals with different costs (in this case, a low-cost, low-quality signal when people don't put in effort and a high-cost, high-quality signal when people put in effort), then people can coordinate with the low-cost and low-quality signals in order to achieve an uninformative equilibrium in the peer prediction world (even though there exists a much better equilibrium in that world, convergence to this in practice requires that enough people do not take the 'easy way out'). Therefore, with this in mind, especially with an eye toward real-world implementation on AMT (Amazon Mechanical Turk), we focused on the setting in which we have limited access to ground truth, which we use to evaluate the effort of voters.

Additionally, there exists literature along the lines of disincentivizing agents with proficiency lower than a desired threshold from participating in the overall mechanism (18), but we do not pursue this approach, as disincentivizing people from participating by giving out negative rewards would err

on the side of crowding out many potential participants, and we would like to ideally ensure that as many people participate as possible in the peer evaluation portion of these mechanisms.

We are also assuming that each agent gets (monetary) utility from both their final rank and their bonus for "accuracy", and not any additional utility from their final rank (e.g., in an online labor market, people may benefit from being near the top of a list, which may outweigh any lack of bonus they incur from not putting in any effort to compare their peers). In particular, in order to measure both an agent's utility of her final rank and the bonus received on the same scale, we map positions in the final ranking to monetary valuations and then combine them with bonuses. We touch upon this more in the following section.

3 Model

We first introduce necessary background notation. As in (9), we let $[k] = \{1, \ldots, k\}$ for any positive integer k. We denote the number of players as n, and the set of all players is $[n] = \{1, \ldots, n\}$. The opinion of each player is represented by a *permutation* over $[n]$, which is a ranking of all n players (including that player herself). We also let Π represent the set of all permutations (possible rankings) of $[n]$ and Π^n represents the set of all *input profiles*, or collection of opinions from the n players. Let $\sigma \in \Pi$ denote a ranking, $\sigma(j)$ denote the player at position j in ranking σ, and $\sigma^{-1}(i)$ denote the position of player i in the ranking σ. For clarity, we think of position 1 as the best (highest) and position n as the worst (lowest).

Let there be n agents, each of whom makes $d = O(n \log n)$ pairwise comparisons of other agents (evaluations), where agent i's pairwise comparisons will be used to generate her opinion of the entire ranking, σ_i. Although agent i does not report a complete ranking σ_i, because agent i reported enough comparisons, we can use standard concentration inequalities to say that this is approximately equivalent to them reporting a complete ranking because even a small number of reports per comparison accurately estimates the pairwise comparison matrix. For more details, please see the discussion in (9). Let each agent i have a true quality q_i that is drawn from some distribution with mean μ_i and some variance (i.e., let these be drawn from a parameter-based model—for example, the Thurstone-Mosteller or Bradley-Terry-Luce models). Also, let $e_i \in \{0, 1\}$ denote the amount of *effort* agent i puts into evaluating her peers; if $e_i = 0$, she randomly flips a coin for every comparison she is asked to make, and if $e_i = 1$, for every comparison she makes, she samples one value from each person's distribution and reports the player with the higher value as being better. In particular, each pairwise comparison is made independently. Putting in effort costs a constant c, whereas putting in no effort costs nothing. Additionally, in this setting, unlike in the previous work on impartiality, agents have valuations for each rank of the final ranking. In particular, we assume linear valuations: each agent values rank j at $(n - j)\delta$, for some positive constant δ.Therefore, the first position in the final rank is worth $(n - 1)\delta$ and the last position in the final rank is worth 0. As an aside, this is somewhat reminiscent of the Borda count, although in an obviously very different context—however, the Borda count does serve as the intuition behind why we assume linear valuations for positions.

Additionally, we assume a two-stage evaluation process for ranking that involves bonus-based rank aggregation rules. Intuitively, a *bonus-based rank aggregation rule* f is a rule that takes an input profile $\vec{\sigma}$, valuations on positions $\vec{v}$, and a bonus payment b and returns a final ranking of agents sorted by total reward. First, f generates a preliminary ranking and assigns each agent a reward corresponding to the value of the position in which that agent was placed. Then, depending on the accuracy of each agent's report, the rule adds a bonus to each person's current reward and then returns a ranking of players sorted in decreasing order of reward. In particular, throughout this paper, the first stage of aggregation will be done by an impartial rank aggregation rule. Note that using a partial aggregation rule will not result in an impartial algorithm, because assigning bonuses takes into account one's own preferences, so if the initial ranking step also takes into account one's own preferences, then no guarantees of impartiality can hold.

The reason for this two-stage evaluation process is because it is very easy to show that if you completely separate the final ranking from all bonuses, it is very easy to ensure both impartiality and effort-incentivization separately (for example, see the results for the low bonus setting in 4.2.1.

Also, we assume that we're in a setting in which we have limited access to ground truth. This is the case in many real-world applications of crowdsourcing: in MOOCs, there are TAs who are trusted to grade problems correctly, and even on hiring platforms, we can hire experts to compare selected

pairs of applicants. This is a common assumption in this literature space; for example, Gao et al. (6) assume a similar limited access to ground truth. Given this ground truth, we will require that each agent make d' comparisons over the ground truth we have. Throughout the rest of the paper, we will assume that $d' = O(\log n)$. Note, however, that although agents make d' comparisons over the available ground truth, they still make $d = O(n \log n)$ comparisons in total.

Now, let us rigorously define impartiality and effort-incentivization. We use the same definition of impartiality as in (9).

Definition 3.1 *A (possibly randomized) rank aggregation rule f is* impartial *if for all $i \in [n]$, all input profiles $(\sigma_1, \ldots, \sigma_n) \in \Pi^n$, and all $\tilde{\sigma}_i \in \Pi$, it holds that $\vec{x} = \vec{y}$, where x_j is the probability i is ranked in position j in $f(\sigma_1, \ldots, \sigma_{i-1}, \sigma_i, \sigma_{i+1}, \ldots, \sigma_n)$, and y_j is the probability i is ranked in position j in $f(\sigma_1, \ldots, \sigma_{i-1}, \tilde{\sigma}_i, \sigma_{i+1}, \ldots, \sigma_n)$.*

Definition 3.2 *A (possibly randomized) rank aggregation rule f is* effort-incentivizing *if for all $i \in [n]$, all input profiles $(\sigma_1, \ldots, \sigma_n) \in \Pi^n$, and all $\tilde{\sigma}_i \in \Pi$, it holds that $\mathbb{E}[g(i_E)] > \mathbb{E}[g(i_N)]$, where i_E is i's report with effort, i_N is a report with no effort, and g is the total reward given to i. Note that a report with no effort exactly corresponds to choosing a ranking at random, which is why we require strict inequality here.*

3.1 Impartial Mechanisms

In order to build intuition about what an impartial mechanism is and what guarantees it provides, we present a high-level overview of two impartial mechanisms in (9).

NAIVE-BIPARTITE: Intuitively, this algorithm randomly partitions the n voters into two subsets of roughly equal size, X and Y. It then uses the opinions of voters in X to get a ranking over voters in Y, as well as vice versa. Then, it deterministically interleaves these rankings to return a final ranking. Note that this is impartial because voter i's position is uniquely determined by i's position in the ranking over voters in her subset, which is determined only by the opinions of voters in the other subset.

COMMITTEE: Intuitively, this algorithm selects a committee of size k at random, which then determines the entire ranking. Let the committee be $X = \{x_1, \ldots, x_k\}$. The algorithm then proceeds in two stages. First, for each committee member x_i, we determine her rank using only the rankings given by the remaining $k - 1$ members. We must do some additional bookkeeping to make sure there are no collisions when placing committee members; for a more detailed explanation (which involves reserving slots in bins for each member), see the paper. After putting committee members in place, there are $n - k$ slots left to fill. In the second step, the committee places the remaining $n - k$ players in the open slots in order in which the committee as a whole ranks them. This is impartial because each committee member is placed in a slot completely determined only by the opinions of other committee members, and each non-committee member has no say whatsoever in this algorithm.

4 Theoretical Contributions

With this definition of the model in mind, we now turn to the problems of bonus-determination and simultaneously ensuring impartiality and incentivizing effort.

4.1 Determining Bonuses

Now, in order to give people bonuses for putting in effort, we define the following payment rule. Let $i, j = arg\ min_{i,j \in [n], i \neq j} |q_i - q_j|$ be the closest two agents i and j in terms of true value. Without loss of generality, assume that agent i has a higher true quality than agent j, and let $p_{min} = \mathbb{P}[i \succ j] = f(q_i - q_j)$. We know that $p_{min} > 1/2$; in particular, we let $p_{min} = 1/2 + \alpha$ for some $\alpha > 0$. Also, every agent who puts in effort will make each comparison correctly with probability at least p_{min}. Therefore, we can set up a payment rule that tries to determine if someone is not putting in effort and essentially flipping random coins ($p = 1/2$) or if someone is actually putting in effort and succeeding on each individual comparison with probability at least p_{min}. We therefore define the following rule.

Thresholded Bonus Determination: Given an input profile $\vec{\sigma}^n$, do the following to determine the distribution of bonuses. First, estimate p_{min} from the data, and make sure it is an underestimate of the true p^*_{min}. Let $p_{min} = 1/2 + \alpha$. Let $\varepsilon = \alpha/2$. Now, for each agent, examine her set of d' ground-truth checkable comparisons. If at least $(p_{min} - \varepsilon)d'$ of these comparisons agree with ground truth, label her as someone who put in effort (E). If fewer than $(p_{min} - \varepsilon)d'$ of these comparisons agree with ground truth, then label her as someone who put in no effort (N). Once you have these classifications, reward the voter with a payoff b if she is classified as an E and with payoff 0 otherwise. This will result in the correct classification with probability at least $\exp\left(-2d'\varepsilon^2\right)$, and as n increases, the probability of error goes to 0.

Proof: Note that the probability that we misclassify an E as an N is the probability that an E answers fewer than $(p_{min} - \varepsilon)d'$ comparisons correctly. The probability that we misclassify an N as an E is the probability that an N answers more than $(p_{min} - \varepsilon)d'$ comparisons correctly. Let X_E be the random variable that denotes the number of correct answers an E would give, and let X_N denote the number of correct answers that an N would give.

First, examining X_E, we can rewrite what we want as

$$\begin{aligned}\mathbb{P}\left[X_E < (p_{min} - \varepsilon)d'\right] &= \mathbb{P}\left[X_E - \mathbb{E}\left[X_E\right] < (p_{min} - \varepsilon)d' - \mathbb{E}\left[X_E\right]\right] \\ &\leq \mathbb{P}\left[X_E - \mathbb{E}\left[X_E\right] < -\varepsilon d'\right]\end{aligned}$$

where the second transition occurs because $\mathbb{E}\left[X_E\right] \geq p_{min}d'$ due to how we defined p_{min}.

Applying Hoeffding's inequality to the above, we can bound the probability of error in each case.

$$\begin{aligned}\mathbb{P}\left[X_E - \mathbb{E}\left[X_E\right] < -\varepsilon d'\right] &\leq \exp\left(\frac{-2(-\varepsilon d')^2}{\sum_{i=1}^{d'}(b_i - a_i)^2}\right) \\ &= \exp\left(\frac{-2(\varepsilon d')^2}{d'}\right) \\ &= \exp\left(-2\varepsilon^2 d'\right).\end{aligned}$$

Now, examining X_N, we have

$$\begin{aligned}\mathbb{P}\left[X_N > (p_{min} - \varepsilon)d'\right] &= \mathbb{P}\left[X_N - \mathbb{E}\left[X_N\right] > (p_{min} - \varepsilon)d' - \mathbb{E}\left[X_N\right]\right] \\ &= \mathbb{P}\left[X_E - \mathbb{E}\left[X_E\right] > (\alpha - \varepsilon)d'\right]\end{aligned}$$

because $\mathbb{E}\left[X_N\right] = d'/2$. Also, remember that we defined $\alpha = 2\varepsilon$, so we can rewrite this and apply Hoeffding again as follows.

$$\begin{aligned}\mathbb{P}\left[X_E - \mathbb{E}\left[X_E\right] > \varepsilon d'\right] &\leq \exp\left(\frac{-2(\varepsilon d')^2}{\sum_{i=1}^{d'}(b_i - a_i)^2}\right) \\ &= \exp\left(\frac{-2(\varepsilon d')^2}{d'}\right) \\ &= \exp\left(-2\varepsilon^2 d'\right).\end{aligned}$$

Note that the probability of error in both cases goes to 0 as n increases because we assume that $d' = O(\log n)$. □

Now that we have a principled way of classifying voters based on perceived effort with high probability as n increases and the associated bonus scheme, let us examine how different sizes of bonuses provide different guarantees of effort-incentivization and impartiality.

4.2 Low and High Bonus Settings

We now introduce two settings: the *high bonus* and *low bonus* settings. Let b be the bonus for putting in effort, and let $c < b$ be the cost of putting in effort. In the high bonus setting, $b \geq \delta$, and in the low bonus setting, $b < \delta$. Intuitively, in the low bonus setting, effort-based bonuses cannot affect peoples' final rankings, whereas in the high bonus setting, the effort-based bonuses are large enough to change the overall ranking.

4.2.1 The Low Bonus Setting

In the low bonus setting, note that rewarding bonuses to voters cannot affect the final ordering because $b < \delta$. That is, even voter i in position k and voter j in position $k-1$ both put in effort, but voter i gets the bonus and voter j does not, i still gets reward $(n-k)\delta + b - c$ and j gets reward $(n-k+1)\delta - c$, meaning $isuccj$ in the final ranking because $b < \delta$. Therefore, in this setting, all previous impartial rules remain impartial when bonuses are added. Furthermore, these rules satisfy effort-incentivization because putting in effort increases the chance that you will be given a bonus, which will increase your total reward. In a sense, in the low bonus setting, it is possible to treat impartiality and effort-incentivization completely separately.

Additionally, all notions of error measures and accuracy guarantees exactly carry over from previous work here because the bonus mechanism has absolutely no bearing on the final ranking returned by the impartial mechanism. This will not be the case in the high bonus setting.

4.2.2 The High Bonus Setting

In the high bonus setting, rewarding bonuses to voters can affect the final ordering. This obviously breaks impartiality because now, investing effort and reporting more accurately can allow agent i at position k to leapfrog agent j at position $k-1$ if agent j does not also get an accuracy bonus for putting in effort. However, this setting allows us to guarantee the following notion of *effort-impartiality*.

Definition 4.1 *A (possibly randomized) rank aggregation rule f is* effort-impartial *if for all $i \in [n]$, all input profiles $(\sigma_1, \ldots, \sigma_n) \in \Pi^n$, and all $\tilde{\sigma}_i \in \Pi$, the following holds for every agent i. If i puts in effort, with high probability this will either not change the distribution of her final location the same or will result in a strictly better distribution of outcomes; and, if i doesn't put in effort, with high probability this will either not change the distribution of her final location or will result in a strictly worse distribution of outcomes. Essentially, effort-impartiality guarantees that putting in effort will, with high probability, never harm you, and that not putting in effort will, with high probability, never help you.*

In order to see that any impartial rule coupled with a high bonus scheme is effort-impartial, we will show that, with high probability, putting in effort will never decrease someone's final position, and that not putting in effort will never increase someone's final position. In the first case, you will be rewarded for putting in effort based on your agreement with the limited ground truth available. By our earlier result, this happens with high probability. Now, note that because you receive a bonus, no one at any spot lower than yours can 'leapfrog' you and end up higher in the final ranking, and you could potentially leapfrog people ahead of you and end up in a better position. In the second case, with high probability you will not be rewarded, and therefore you cannot leapfrog anyone ahead of you in the ranking. In fact, you may be leapfrogged by people below you in the ranking, so you cannot increase your final position and may be penalized for not putting in effort.

Note that many of these results somewhat separate the effects of impartiality and effort-incentivization—over the course of the project, we found that it was very hard to look at them together, and it seems like they are somewhat at theoretical odds with each other. The concept of impartiality is that your report can't affect your final rank, and the concept of effort-incentivization is exactly the opposite—that your report can affect a reward you receive—so putting them together in a cohesive manner is quite difficult.

Also, note that error guarantees are much harder in the high bonus setting. This is mostly due to the fact that, without restrictions on the size of bonuses, it could be possible for a voter ranked last by the impartial mechanism to end up in the first position if the potential bonus is large and she is the only one who puts in effort while doing comparisons.

It also seems hard to leverage effort-incentivization to get potentially better error guarantees in general. One interesting variant of an impartial algorithm that we initially thought would not only incentivize people to put effort into their comparisons but could perhaps lead to better accuracy guarantees (i.e., higher-quality reports lead to more accurate aggregate rankings) is the following variant of the COMMITTEE algorithm in (9). The bulk of the algorithm stays the same, but the way in which we choose the committee is a bit different. Instead of randomly selecting the committee, sort the voters based on agreement with ground truth. In particular, based on our bonus payment rule, the top k' will be paid a bonus and the bottom $n - k'$ will not. However, this rule is not quite impartial

because now your report can affect who is on the committee. In particular, your report can affect whether you're on the committee or not, which will then change the way in which you are placed in a slot. This points out the difficulty inherent in trying to use accuracy and effort to pre-process data for an impartial rule; all of these 'smoothing' steps break impartiality because it breaks symmetry among the voters.

5 Empirical Understanding of Effort-Impartial Setting on AMT (N=90)

Crowdsourcing platforms such as Amazon Mechanical Turk (AMT) bring together requesters (employers) and workers (employees), where requesters post open calls for workers to complete short, micro tasks in return for monetary compensation (12). With access to over 15K active "Master Turkers", Turk Workers who have received the Master's qualifications for their Turking expertise, such a platform can be used as an ecologically valid experimental environment (14). We therefore leverage this platform to deploy our effort-impartial algorithms.

To date, there is little understanding of how crowd workers understand and by extension behave based on different scoring rules, differing incentivization schemes, and how such participant understanding and behavior evolves over the long term. Therefore, to inform research questions and experimental design, we connect the dots between recent theoretical crowdsourcing work, as well as social psychology's understanding of human decision making and perception of algorithms. For instance, work in progress by (10) found that leveraging a behavioral-based framing (13) of an impartial mechanism led to a 30% decrease of dishonest behavior, as compared to no mention of such impartial mechanism, a policing framing (4), and self-concept framing (7). However, such work found that by communicating the presence of the impartial algorithms to participants through a behavioral framing, a seemingly inadvertent effect of occured: participants perceived diminishing returns for effort invested in the task. We therefore iterate on the framing and incentive structure of the impartial setting, by incorporating an effort-bonus schema as constructed in sections 4.3.1 and 4.3.2.

5.0.1 Methods and Experimental Design

In an AMT between-subjects experiment (N=90), participants, i.e. Master Turkers, were in one of three conditions: control (i.e. no mention of effort-impartiality, N=15), small payment bonus (N=15), and big payment bonus (N=15, explained in detail below). We also collect full ranking from veteran Master Turkers: those who have completed over 10,000 Human Intelligence (N=45, 15 per condition). While sample size is too small to make statistical conclusions to whether differences exists across conditions, we are able to better understand a first pass of deploying such effort-impartial algorithms in the real world.

We maintained the simple two-step task (10), which leverages Master Turkers domain expertise (i.e. Turking on AMT), as well as ensures the set of voters is the set of alternatives. In the first step (20 min), Master Turkers generated one to two paragraphs of advice for new Amazon Mechanical Turk workers. In the second step (20 min), these same Master Turkers complete at least N pairwise comparisons of other Turker's advice. Base rate compensation for both parts was $10/hr and participants had opportunity to collect a bonus if their advice was ranked overall in position 1: $5, Rank 2: $4 and so on (i.e. incentivizing to be ranked in top k overall). The behavioral framing is as follows: "Remember, the comparisons you generate will have no effect on where your advice is ranked in the final ranking through an impartial mechanism" (10). At the end of this string, we added the two differing effort-bonus schemas depending on condition, either: 1. Small payment: "You will receive an additional bonus for each accurate comparison you make: $0.05 for every correct comparison as compared to expert ground truth for potential total of $1.00" and 2. Big payment: "You will receive an additional bonus for each accurate comparison you make: $0.10 for every correct comparison as compared to expert ground truth for potential total of $2.00" – a bonus that is just shy of total base-rate compensation, and therefore relatively large.

To better understand worker perceptions of effort, fairness, and accuracy, post-task survey questions were completed at the of Part 2. For example: "I put in a lot of effort in Part 2 of this study: comparing my peers' advice," "My peers put in a lot of effort in Part 2 of this study: comparing my peers' advice," "The amount of effort I put in Part 2 (advice comparison) affects my own position in the final ranking." Questions for fairness, honesty, and accuracy were also asked in a similar format. Responses were 5-point Likert scale, from Strongly Agree to Strongly Disagree.

To extend prior work, our first research question is: **RQ1** in an effort-impartial setting, do differing payment schemas lead to differing amount of invested effort in task completion? We hypothesize **H1** that largest effort payment schemas leads to highest effort invested. Our second research question is: **RQ2** in an effort-impartial setting, how does effort invested change accuracy? We hypothesize **H2** that the more effort invested, the more accurate pairwise comparisons will be (compared to expert ground truth) as participants are task-domain experts. And lastly, our third research question: **RQ3** in an effort-impartial setting, do participants perceive effort investment differently across conditions, both of themselves and their peers? We hypothesize **H3** that participants perceive highest effort invested in largest payment schema.

5.0.2 Results

Our findings suggest a two-fold consequence of introducing payments: 1. both perceived effort and time on task increase when effort payment is highest and 2. increased effort does not lead to increased accuracy, as compared to expert ground truth.

We first leverage duration of task as somewhat noisy proxy for effort. In the control condition, average time spent completing the task is 53 seconds/comparisons, and in the small-payment condition, average duration is 42 seconds/comparisons. In the big-payment condition average duration is almost double: 81 seconds/comparisons. This supports, although rather noisily, **R1**: workers invest more effort (measured through duration on task) when effort is rewarded independently of rank aggregation reward.

Furthermore, we see that across all three conditions, the "quality" of expert data remains seemingly consistent. Here we measure quality through agreements: the ratio of the sum of majorities to the maximum possible (everyone is unanimous). In the control condition, experts have 87.5% agreement; in the small-payment schema, experts have 87.8% agreement; and in the big-payment schema, experts have 86.6% agreement. We can leverage this relatively stable measure to then better understand how agreement within voters compares across conditions. We find that, as expected, the amount of voter agreement seems to increase by introducing effort bonuses: 75.4%, 93.5%, 83.4%. However, surprisingly, we find that such agreement amoungst voters (who, again, are domain experts) does not lead to increase in accuracy, as compared to ground truth. For instance, looking at the Committee aggregation (other mechanisms follow a similar pattern) across conditions there is the highest amount of accuracy in the control condition (77.5% recall), and recall decreases as effort-bonus increases: 67.3% recall in small-payment condition and 58.8% recall in big-payment condition. Therefore, **R2** we find a seemingly opposite result from our hypothesis. We can then turn to the qualitative data collected to shed light on such unexpected results. We find that across the small-payment and big-payment conditions, participants perceive fairness, accuracy, and honesty similarly. However, **R3** in the big-payment schema, participants are more likley to report that they invested a lot of effort, as compared to small-payment condition ($p<0.05$, Wilcoxon Rank Sum Test). That is, participants were more likely to agree to the following Likert-scale question: "I put in a lot of effort in Part 2 of this study: comparing my peers' advice." Such results point to a need for future work to investigate mental models of voters: why is it that when there is no mention of impartiality, effort is perceived to play a role in final ranking? How can we fine tune this task to leverage domain expertise of voters as robust measure for accuracy?

6 Discussion and Future Work

On the theoretical side, this problem turned out to be slightly underwhelming. It turns out that in at least the setting with limited access to ground truth, the problem nicely splits into two settings: one in which bonuses have the potential to change the initial ranking, and one in which bonuses do not. In both cases, essentially the best you can reasonably do is by combining previous impartial rules with a thresholded bonus mechanism that tries to identify high- and low-effort participants. Therefore, people are essentially *independently* incentivized to put in effort irrespective of their final rank because of the bonus associated with effort. However, this is not particularly interesting because it doesn't seem to break much additional theoretical ground on the two of them together. Also, accuracy guarantees remain exactly the same in the low bonus setting, but can conceivably be arbitrarily bad in the high bonus setting, depending on the ratio between bonuses and positional values.

There also does not yet exist a principled way of relating the size of the bonus and the positional values in the final ranking, and it would be nice to flesh out this tradeoff more clearly. Perhaps, by weakening impartiality slightly, but not allowing arbitrarily large bonus payoffs, we can actually strengthen accuracy guarantees because people will be more willing to report correctly for a large bonus. However, in the extreme where $b >> n\delta$, then all that the participants will care about is getting the bonus, so presumably they will all put in effort, resulting in potentially much better data and therefore accuracy guarantees conditioned on their expected behavior given this bonus (note that when all participants are classified as having put in effort, impartiality is preserved).

We thank Professor Nihar Shah for his feedback on this work, as well as our advisors Ariel Procaccia and Chinmay Kulkarni. This work was conducted under IRB Protocol STUDY2017_00000224.

References

[upw] Online work report: Global, 2014 full year data. Technical report, Upwork.

[2] Agrawal, A., Horton, J., Lacetera, N., and Lyons, E. (2013). Digitization and the contract labor market: A research agenda. Technical report, National Bureau of Economic Research.

[3] Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

[4] Allingham, M. G. and Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of public economics*, 1(3-4):323–338.

[5] Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 193–202, New York, NY, USA. ACM.

[6] Gao, A., Wright, J. R., and Leyton-Brown, K. (2016). Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse. *CoRR*, abs/1606.07042.

[7] Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American psychologist*, 35(7):603.

[8] Horton, J. J. and Golden, J. M. (2015). Reputation inflation: Evidence from an online labor market.

[9] Kahng, A., Kotturi, Y., Kulkarni, C., Kurokawa, D., and Procaccia, A. (2018). Ranking wily people who rank each other. To appear in AAAI 2018.

[10] Kotturi, Y., Kahng, A., Kurokawa, D., Procaccia, A., and Kulkarni, C. (2018). Hirepeer: Impartial peer assessment accurately identifies qualified crowd workers despite conflicts. Work In Progress.

[11] Kuek, S. C., Paradi-Guilford, C., Fayomi, T., Imaizumi, S., Ipeirotis, P., Pina, P., and Singh, M. (2015). The global opportunity in online outsourcing. World Bank Other Operational Studies 22284, The World Bank.

[12] Mason, W. and Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23.

[13] Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644.

[14] Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk.

[15] Prelec, D. (2004). A bayesian truth serum for subjective data. *science*, 306(5695):462–466.

[16] Shah, N. B. and Zhou, D. (2015). Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In *Advances in neural information processing systems*, pages 1–9.

[17] Silberman, M., Ross, J., Irani, L., and Tomlinson, B. (2010). Sellers' problems in human computation markets. In *Proceedings of the acm sigkdd workshop on human computation*, pages 18–21. ACM.

[18] Witkowski, J., Bachrach, Y., Key, P., and Parkes, D. C. (2013). Dwelling on the negative: Incentivizing effort in peer prediction. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[19] Witkowski, J. and Parkes, D. C. (2012). A robust bayesian truth serum for small populations.

7 Appendix

7.1 Free Lunch

One early result we hoped to establish was an analogous *no-free-lunch (NFL)* axiom, as in (16). However, it quickly became clear that this is impossible when dealing with full impartiality. (Bounded size of bonuses...how to evaluate everything? Just frame it as 0 payoff if not putting in effort; but gets placed somewhere else that doesn't depend on his own opinion.)

The no-free-lunch theorem as presented in (16) states that no one can get a positive payoff if they skip all the questions asked in a crowdsourcing setting. We slightly modify the statement to fit our own setting; in particular, there is no option to skip comparisons; we just want people to have to put in effort. Therefore, our *impartial-no-free-lunch axiom* states that putting in no effort when evaluating comparisons results in an overall payoff of 0. However, note that this is quite patently false: by definition, an impartial mechanism ensures that agent i's report cannot affect her final position, and therefore if the person ranked in the top position puts in no effort, this doesn't affect the fact that she is placed in the top spot overall and receives nonzero utility for her placement there. Therefore, impartiality and no-free-lunch results are quite obviously and forever at odds.