Decentralized Oracles via Peer-Prediction in the Presence of Lying Incentives

Naman Goel and Aris Filos-Ratsikas and Boi Faltings
Artificial Intelligence Laboratory
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland, 1015
{naman.goel, aris.filosratsikas, boi.faltings}@epfl.ch

Abstract

We derive conditions under which a detail-free minimal peer prediction mechanism can be used to elicit truthful data from non-trusted rational agents when an aggregate statistic of the collected data affects the amount of their incentives to lie. Furthermore, we discuss the relative saving that can be achieved by the mechanism, compared to the rational outcome, if no such mechanism was implemented. Our work is motivated by distributed platforms, where decentralized data oracles collect information about real-world events, based on the aggregate information provided by often self-interested participants. We compare our theoretical observations with numerical simulations on two publicly available real datasets.

1 Introduction

With the increasing popularity of the blockchain technology in recent years, the implementation of commercial and governmental systems has witnessed a large shift towards distributed and decentralized approaches. In particular, the emergence of the Ethereum platform has given rise to the development of several applications (often referred to as *decentralized apps* or *DAPs*) which aim to apply this latter principle to many areas of interest such as finance, education, intellectual property or government.[1] At the heart of these approaches lies the concept of the *smart contract*, i.e., lines of code that contain the terms of the agreement between the involved parties, which are automatically executed once triggered by events happening in the real world. For example, consider the case of a web service, which is typically dictated by a service level agreement (SLA) between the service provider and the clients. The SLA can be coded into a smart contract between the involved parties which will trigger an automatic payment upon detection of a violation. For instance, if the service guarantees a `responseTime` of at most 1 second with a high probability, frequent slower responses would result in the clients being compensated by the smart contract.

An important issue here is, how do we determine whether the real-world event has actually happened? In the words of Ari Juels, co-director of the Initiative for Cryptocurrencies and Contracts (IC3)[2], as quoted in (Peck 2017): *"Anything they (blockchains) learn about the outside world has to be injected into them"*. The services responsible for acquiring and validating such data about real-world events are called *oracles* and often appear in the form of separate entities; in the web service example, these could be third-party companies that monitor and report traffic. This approach however is prone to several problems, such as the trustworthiness or the accuracy of the oracles, or the cost of maintaining such external services, and in a sense is in conflict with the decentralized nature of the blockchain technology.

An alternative solution would be to appeal to the "wisdom of the crowds" and gather the information about the real world from the participants of the platform themselves. This is the fundamental idea behind the prediction-market applications that run on the Ethereum blockchain, such as Augur (Peterson et al. 2018) or Gnosis (Gnosis 2017). Another concrete example of this principle is the Decibel.Live application (Kandy and Loomb 2017), for compensations due to excessive noise levels from nearby construction companies. In Decibel.LIVE, the client signs a smart contract with the company, specifying the accepted levels of noise in a neighborhood, and any noise level exceeding those thresholds triggers an automatic compensation payment. The noise levels are measured by noise monitoring sensors (installed locally, perhaps in or near the houses of the clients) and are reported via a smartphone app to the platform; this is largely beneficial in that it does not require a trusted authority for the monitoring and the collection of the data.[3]

A genuine concern that typically accompanies this type of decentralized approaches is whether people can be trusted to provide correct information. This concern is even more amplified in settings where people have *rational incentives* to provide false information. In the example above, it seems reasonable that clients might attempt to fake the noise levels in order to get compensated (since generating noise is generally an easy task) or, in the web service example, the clients would have an incentive to always report high response times, in order for the conditions of the smart con-

[1] https://www.stateofthedapps.com

[2] http://www.initc3.org/

[3] The platform does offer the possibility of hiring trusted authorities for monitoring as well, but the smartphone app is advocated as a better and more cost-effective option.

tract to be violated in their favor. We refer to such incentives as *outside incentives* and we will be interested in problems of this specific nature.

It is fairly obvious that, if no additional measures are taken, the rational outcome would be one in which the participants get compensated, regardless of whether there was truly a violation of the terms of the smart contract or not. To counteract this phenomenon, the agents need to be properly incentivized by the platform to provide their feedback truthfully. One way to do this is to issue them a side-payment in addition to the payment that they receive based on the outcome resolution. The side-payment scheme has to satisfy the following two desiderata, (a) to ensure that it extracts the correct information from the participants and (b) to achieve as much saving as possible, compared to the compensation of the rational outcome. In a nutshell, we are interested in the following question:

When the agents are rational, how can the platform determine the correct outcome by acquiring correct information while ensuring as much saving in payments as possible?

For a side-payment scheme to achieve the goal set in (a), clearly the side-payments have to be contingent on the truthfulness of the agents' reports. However, since we do not have access to a ground truth and therefore, there is no way to directly establish the truthfulness of the feedback, we will appeal to the power of *peer-prediction mechanisms* (Miller, Resnick, and Zeckhauser 2005) to align the incentives of the agents.

Truthful feedback elicitation and challenges

Truthful feedback elicitation is a well-studied problem in the literature. Elegant game-theoretic mechanisms, called *peer-prediction mechanisms* (Miller, Resnick, and Zeckhauser 2005; Prelec 2004), exist for eliciting truthful feedback from rational, self-interested agents. Generally speaking, the idea in those mechanisms is to match the report of an agent with that of a randomly chosen peer, and provide a payment as a function of the relation between the two reports. Recently, there has been a lot of progress towards making the peer-prediction mechanisms useful for practical applications. The modern peer-prediction mechanisms are generally *minimal*, i.e., they require the agents to submit only their feedback report and *detail-free*, i.e., they do not require any knowledge of the agents' beliefs. Several mechanisms in this class have been designed, e.g., see Dasgupta and Ghosh (2013), Shnayder et al. (2016), Radanovic and Faltings (2015), Kamble et al. (2015); mostly useful to us will be the *Peer Truth Serum for Crowdsourcing* mechanism of Radanovic, Faltings, and Jurca (2016) (henceforth PTSC).

Peer-prediction mechanisms are known to incentivize agents to invest the effort (which can be interpreted as a cost) required to make observations and to subsequently report their observations truthfully. These mechanisms have been shown to counter two types of lying incentives: one is the potential saving in the cost of effort (by not making any observations, but rather submitting uninformed reports) and the other is the *internal* reward, since rewards in a non-truthful incentive mechanism may also create unintended incentives to lie.

However, the truth-telling properties of minimal and detail-free mechanisms has not yet been studied in settings when agents have outside incentives to lie, like the settings mentioned in the introduction, where the agents receive an outside reward as a function of the aggregation of their reports. Since every agent's feedback has a "share" in determining the outcome, the agents have variable lying incentives that also *depend on the strategies adopted by other agents*, which makes the "overcoming of extra incentives" more challenging.

Our contributions

In this paper, we consider settings with outside incentives and binary observations, and we employ the PTSC mechanism of (Radanovic, Faltings, and Jurca 2016) as a side-payment scheme. We prove that, with an appropriate choice of the scaling constant, the mechanism can be used to ensure that truth-telling is a strict equilibrium of the induced game. Furthermore, we show that if there is *any positive* fraction f of *honest* agents (i.e., agents that always report truthfully), the strategy profile in which the agents exercise their outside incentives, or *denial strategies* (e.g., reporting "bad" service) is no longer an equilibrium. We note that assuming the *existence* of honest agents is very different from using trusted authorities since our method does not depend on knowing who these honest agents are. This is a rather common scenario, as in a large platform, one would normally expect at least a few agents to behave honestly but we would not expect to know their identities. These properties of the PTSC mechanism were already known in the *absence* of the outside lying incentives (Radanovic, Faltings, and Jurca 2016); our paper extends the analysis of PTSC and establishes its truth-telling incentive properties in the *presence* of the outside lying incentives.

Additionally, for the first time, we compute a bound on the scaling required for ensuring a truth-telling equilibrium of the side-payment scheme, as a function of the outside incentives. We also provide conditions under which the side-payment scheme gives positive saving compared to the rational outcome (i.e., the denial strategy outcome) and we prove a lower bound on this saving. We show that as the number of agents grows large, the saving approaches the best possible saving, attainable when all agents are honest, without any side-payments. We also provide bounds (on the same quantities) when the scaling has to not only ensure a truthful equilibrium, but also to eliminate the denial strategy equilibrium. Interestingly, in the process of doing this, we find an upper bound on the fraction of honest agents that should be present, in order for the side-payment scheme to still be profitable.

Finally, the scaling constant, as well as the savings of PTSC depend on a quantity δ^*, which we refer to as the *self-predictor value* and is essentially a measure of correlation strength between prior and posterior signals. The assumption that $\delta^* > 0$ is a standard assumption in the literature of peer-prediction (e.g. see (Jurca and Faltings 2005; Witkowski and Parkes 2012a)) and translates to positive cor-

relation between the observations of the agents. We quantify the required scaling constant as well as the saving in terms of this quantity. Moreover, we do not need to know this quantity; an *estimate* is sufficient for the results to either hold exactly or *approximately*, where the approximation error goes to 0 as the number of agents grows large.

Related Work

Our work draws on the recent ideas in the peer-prediction literature (Miller, Resnick, and Zeckhauser 2005; Prelec 2004; Witkowski and Parkes 2012b; Radanovic and Faltings 2013; Dasgupta and Ghosh 2013; Waggoner and Chen 2014; Radanovic and Faltings 2015; Kamble et al. 2015; Shnayder et al. 2016; Radanovic, Faltings, and Jurca 2016; Gao, Wright, and Leyton-Brown 2016; De Alfaro, Shavlovsky, and Polychronopoulos 2016; Agarwal et al. 2017; Liu and Chen 2017a; Kong and Schoenebeck 2018; 2018; Goel and Faltings 2018); here we focus on the results related to settings with outside incentives. A survey of the techniques in this area can be found in (Faltings and Radanovic 2017).

The topic of outside incentives in decentralized platforms has been recently explored in the context of prediction markets, drawing motivation from applications like Augur and Gnosis. Chakraborty and Das (2016) perform equilibrium analysis when the market participants may significantly influence the actual realization of the outcome, in a game which is played in two stages; first the agents trade in the market and then they vote on the outcome. Their model captures the empirical observations in prior work (Chakraborty et al. 2013). These works however only analyze the effects of rational behaviour, rather than aim to counteract it, by implementing appropriate mechanisms. Chen et al.(2011) consider similar two-stage models of prediction markets, where the agents strategize only in the first stage to manipulate the market prices used for the predictions. The authors analyze information aggregation properties of the market and don't consider outcome manipulation, and their setting is thus quite different from ours.

Freeman, Lahaie, and Pennock (2017) study a related setting, where they assume that agents trade honestly in the first stage and only behave strategically in the second. They use a peer-prediction mechanism to elicit truthful votes in the equilibrium of the second stage, and show that under certain conditions, the fees charged by the market are enough to cover the side payments. Interestingly, they also use a similar measure of signal correlation, which they refer to as the "update strength", and they express some of their results using this quantity.

Our work differs from Freeman, Lahaie, and Pennock(2017) in two key aspects. First, our informational assumptions are weaker. In particular, we only require access to a measure of signal correlation (the self-predictor value) and actually, only an estimate of that measure is sufficient. In contrast, Freeman, Lahaie, and Pennock (2017) use the prior distribution of the agents' beliefs, which they obtain from the closing price of the market, enabled by the assumption that the agents are honest in the trading stage. While this may be meaningful in a prediction market domain, such assumptions are far less realistic in the more general settings that we consider. Secondly, Freeman, Lahaie, and Pennock (2017) do not address the issue of non-truthful equilibria in their work, which we do through the existence of honest agents. In settings other than prediction markets, Jurca, Faltings, and Binder(Jurca, Faltings, and Binder 2007; Jurca and Faltings 2006) consider feedback elicitation settings in the presence of outside incentives but again, they assume full knowledge of the agents' beliefs.

2 Model and objectives

We consider settings in which a large number of questions are to be resolved on a decentralized platform through acquiring feedback from a finite number of agents per question. The questions can be, for example, of the following form : "Is the `responseTime` of web service W less than 10 seconds?". An agent i makes a private binary observation $X_i \in \{0, 1\}$ about exactly one question and submits her feedback report $Y_i \in \{0, 1\}$ to the platform. For every question, n agents are asked to submit their feedback, and based on this feedback, the questions are said to be resolved by announcing their outcomes. The *outcome* o_w for a question w is defined as the fraction of agents who reported 0 as their feedback. In the web service example, this corresponds to the fraction of agents who report that the `responseTime` of the service was less than 10 seconds.

Note that we define the outcome o_w to be a continuous variable, whereas the feedback is elicited as a discrete variable. This is because of the noisy (and in some cases subjective) nature of the feedback. In the web-service case, `responseTime` is a noisy measurement and no service can promise a certain response time 100% of the time. Thus, it is important to define the outcome as a continuous variable measuring the fraction of time that the service did provide a good response time. The same definition was used in (Freeman, Lahaie, and Pennock 2017).

Based on the announced outcome, every agent (who answered that question) is issued a payment proportional to the value of the outcome. More precisely, the payment given to an agent is $\mathcal{R} \cdot o_w$, where $\mathcal{R}$ is a positive constant. In the web service example, the reason for such payments is the legal contracts that are signed between the web service provider and the agents. Such contracts bind the service providers to issue a refund to their customers if their service does not meet the promised standards.

After making her private observation, agent i uses a strategy σ_i to submit a report Y_i based on observation X_i, in order to maximize her expected payment. The agents are assumed to be *rational* and therefore they will not typically report their true observations, if not properly incentivized to do so. We also follow the common assumption that agents are *risk-neutral*.

Definition 1 (Agent Strategy σ_i)**.** *An agent i's strategy, denoted by $\sigma_i(Y_i = y|X_i = x), \forall\, x, y \in \{0, 1\}$, is the probability of the agent's report for the question being y given that her observation is x.*

The strategy models a variety of possibilities that are available to the agent for mapping her observation to her report. Some examples of such strategies are the following.

Definition 2 (Truth-telling Strategy)**.** *An agent's strategy is called* truth-telling *if and only if* $\sigma_i(Y_i = y|X_i = x) = 1, \forall\, x = y$ *and* $\sigma_i(Y_i = y|X_i = x) = 0, \forall\, x \neq y$.

In *heuristic strategies*, the report of the agents are independent of their observations. One heuristic strategy of particular importance is always reporting 0, formally defined below.

Definition 3 (Denial Strategy)**.** *An agent's strategy is called the* denial strategy *if and only if* $\sigma_i(Y_i = 0|X_i = x) = 1$ *and* $\sigma_i(Y_i = 1|X_i = x) = 0$.

The denial strategy is an interesting strategy in our setting because the payment that agents receive depends on how many of them report 0 as their feedback. The following observation is fairly easy to see, but for completeness, we provide a proof in an online supplement [4]

Observation 1. *In the settings described above, the denial strategy is the strictly dominant strategy for all agents and gives the maximum payment* $\mathcal{R}$.

A strategy σ_i is called strictly *dominant* if it gives agent i her the highest possible payment, given any strategies of the remaining agents. Observation 1 implies that in the presence of rational agents, the outcome determined by the decentralized platform is bound to be 1.00, since every such agent will report 0 irrespective of their true observation. Such an outcome determination is not useful for practical purposes; on one hand, it is not informative and hence provides no utility in terms of the information acquired, and on the other hand, it can incur a huge loss to the platform in practice.

Peer-prediction: To counteract this phenomenon, the agents need to be properly incentivized by the platform to provide their feedback truthfully. We propose to do this, by issuing them a side-payment in addition to the payment that they receive based on the outcome resolution. Clearly, any constant amount of such side-payment does not achieve this objective; the side-payments have to be contingent on the truthfulness of the agents' reports. However, since there is no way to directly establish the truthfulness of the feedback, we will appeal to the power of *peer-prediction mechanisms* (Miller, Resnick, and Zeckhauser 2005) to align the incentives of the agents with their feedback. The most important constituents of the peer-prediction framework are the agents' beliefs about the observations of their peers. We will let $P_i(X_p = x')$, for $x' \in \{0, 1\}$, denote agent i's (prior) belief about a randomly selected peer p's observation X_p on a question being x'. We will assume that all questions are a priori similar so the prior belief of the agent is same for all questions.[5] After the agent makes a private observation X_i for a question, she updates her belief (posterior) about her peer's observation on that question only, to $P_i(X_p = x'|X_i = x)$.

The first objective of this paper is to ensure that the decentralized platform can be used as an oracle, in the sense that the outcome determined by the platform is correct. The next question is, how large do the side-payments need to be? Is it possible to implement the side-payment scheme suggested by the peer-prediction mechanism without incurring loss to the platform? Our benchmark here is the amount of money that the platform would have to pay if there were no side-payments in place, and therefore the outcome would be determined by the denial strategies of the agents. In other words, we define the *relative saving* of a side-payment scheme to be

$$\text{relative saving: } \frac{n\mathcal{R} - \mathcal{P}}{n\mathcal{R}},$$

where $\mathcal{P}$ is the total payment (side-payment + outcome dependent payment) under the scheme to the agents. The reason for considering relative saving in this paper and not the actual saving in monetary units is that the absolute saving is domain and scale dependent and not very informative in a general sense. Before we proceed, let us see what the best relative saving that we could hope for is. The proof of the proposition is deferred to the online supplement.

Proposition 1. *If agents were honest (i.e. they reported truthfully ignoring the outcome dependent payments), the platform could make an expected relative saving of up to* $P(1)$ *in the payments, where* $P(1)$ *is the actual probability of a randomly selected report on the platform being* 1.

Note that the best possible saving is not 100%, because it depends on the actual quality of the service. In the web service example, Proposition 1 states that when the response times of the services are generally good i.e., $P(1)$ is high, the platform could make significant savings (up to 100% as $P(1) \to 1$) if the agents were honest. Also, note that we are comparing against the ideal outcome, when agents would not need to be incentivized to act truthfully; a mechanism that fairs well against this outcome, will fair well against any other side-payment scheme, including one in which the outcome determination is done by a costly third party.

Remark: A slightly different way to quantify the saving is in terms of whether the revenue generated by the service subscription fees is enough to cover the payments. To avoid introducing too many new terms, we postpone the exposition of these results until Section 7.

The PTSC Mechanism. Since we are interested in relaxing the informational assumptions as much as possible, we will use a detail-free mechanism for determining the side-payments on the decentralized platform. In particular, we will use the PTSC mechanism (Radanovic, Faltings, and Jurca 2016), which we describe here for completeness. To decide the reward for an agent, the mechanism selects another agent p who also submitted feedback for the same question. Suppose that the agent submits $Y_i = y$ and the peer submits $Y_p = y'$. The side-payment of $\tau(y, y')$ agent i under the PTSC mechanism is:

$$\tau_(y, y') = \begin{cases} \alpha \cdot \left(\frac{\mathbb{1}_{y=y'}}{R_i(y)} - 1\right) & \text{if } R_i(y) \neq 0 \\ 0 & \text{if } R_i(y) = 0 \end{cases}$$

[4] Due to the page limit, proofs are provided in an online supplement, which is available at http://bit.ly/2BkXjSx.

[5] If not all questions are a priori similar but there are known batches of a priori similar questions, our results can be extended for each batch separately. For example, in the web-services case, this can be done by grouping web-services with similar SLAs

where α is a strictly positive scaling constant. The mechanism uses $R_i(y) = num_i(y)/\sum_{\bar{y}\in\{0,1\}} num_i(\bar{y})$, where $num_i(y)$ is a function that counts occurrences of y in the feedback of all agents (except i) across all questions.

Subjective equilibrium: When referring to the "correct outcome" for rational agents, one needs to define an appropriate *solution concept* is which the outcome will be obtained. The standard objective in the peer-prediction literature is to ensure that the correct outcome is achieved in the equilibrium, or, in other words, that truth-telling is an equilibrium. A strategy profile $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$, which represents a collection of strategies of agents $\{1, 2, \ldots, n\}$, is a *strict equilibrium* if for any agent $i \in \{1, 2, \ldots, n\}$, the agent's expected payment is strictly maximized when she adopts strategy σ_i , i.e. σ_i is her *best response* to the strategies of the other agents. A strategy profile $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$, is an ε*-approximate equilibrium* if for any agent $i \in \{1, 2, \ldots, n\}$, the agent's expected payment when she adopts strategy σ_i, is smaller than the expected payment when she adopts any other strategy σ_i' by at most ε.

Since beliefs need not be common among workers, i.e. they are subjective, the appropriate equilibrium concept that we adopt is the *ex-post subjective equilibrium* (Witkowski and Parkes 2012a), defined over admissible belief types. In this equilibrium concept, a worker's best response is independent of the beliefs of other workers. In the rest of the paper, we simply use the terms "equilibrium" and "ε-approximate equilibrium" for brevity.

3 Truthful Equilibrium and savings

We first derive the conditions under which the PTSC mechanism can be used to ensure that the truth-telling strategy profile is an equilibrium, in the presence of outcome-dependent lying incentives for the agents. This is certainly a critical requirement for a side-payment scheme which elicits reliable information. In the next section, we will provide an even stronger guarantee, ensuring that truth-telling is also a "good" equilibrium, under some reasonable assumptions. In our analysis, we will use the following quantity.

Definition 4 (Self-Predictor Value)**.**

$$\delta^* = \min_i \Big(\frac{P_i(X_p = 1|X_i = 1)}{P_i(X_p = 1)} - \frac{P_i(X_p = 0|X_i = 1)}{P_i(X_p = 0)} \Big)$$

We note that $\delta^* > 0$, whenever the observations of agents are *positively correlated*; this means that conditional on observing 1, the posterior belief of the agent about her peer also observing 1 strictly increases compared to her prior belief about the same. This positive correlation of signals is a standard assumption in the literature of peer-prediction for binary answer spaces, e.g. see (Jurca and Faltings 2005; Witkowski and Parkes 2012a) and it is under this condition that PTSC guarantees that truth-telling is an equilibrium.[6] We will make the same assumption throughout this paper, and we will quantify the required scaling constant of PTSC as well as the relative savings of the mechanism in terms of δ^*. Intuitively, δ^* is a measure of correlation strength, and captures the relative increase in the posterior compared to the prior belief, as described above. A very similar quantity was defined in (Radanovic, Faltings, and Jurca 2016) capturing the very same concept, differing on the fact that it was a multiplicative parameter rather than an additive one. The parameter is also closely related to the *update strength* used in (Freeman, Lahaie, and Pennock 2017).

We emphasize here that the mechanism does not need to know the exact value of δ^*, but we assume that an estimate of this value ($\delta = \delta^* + \beta$, for some $\beta \in \mathbb{R}$) is known. We defer the reader to (Radanovic, Faltings, and Jurca 2016) or (Liu and Chen 2017b), where how a similar estimation could be done is discussed.

Theorem 1. *Given δ and a scaling constant $\alpha > \frac{\mathcal{R}}{n\cdot\delta}$, the truth-telling strategy profile is a strict equilibrium if $\beta \leq 0$, and is a $(\frac{\beta\cdot\mathcal{R}}{n\cdot\delta})$-approximate equilibrium if $\beta > 0$.*

Note that the theorem is stated in terms of ε-approximate equilibria. This is because if the value of δ^* is *overestimated* (i.e., $\beta > 0$), then the agents might have incentive to actually deviate from their truth-telling strategy, but that incentive is bounded by a typically small quantity. In fact, when the overestimation imprecision tends to be negligible (i.e., $\beta \to 0$) or when the number of agents grows large (i.e., $n \to \infty$), then ε goes to 0 and we obtain exact equilibrium. On the other hand, if we only *underestimate* δ^* (i.e., $\beta < 0$), then we obtain exact equilibrium, regardless of the imprecision parameter or the number of agents.

Any overestimation of δ^* does not hurt the saving compared to the case of a precise estimation; in fact, it actually improves it. In contrast, underestimating δ^* can diminish the saving, but the loss again vanishes as the number of agents grows large. The relative savings of the mechanism are captured in the following theorem.

Theorem 2. *The expected relative saving in payments made in the truth-telling equilibrium is at least $P(1) - \frac{1}{n\cdot\delta}$, where $P(1)$ is the actual probability of a randomly selected report being 1 in the truth-telling equilibrium.*

The proof of the theorem is included in Section 2 of the supplement. Note that as long as the condition $n > \frac{1}{P(1)\cdot\delta}$ is satisfied, the lower bound on saving is actually a positive number. Finally, notice that as $n \to \infty$, the relative saving reaches the maximum achievable value $P(1)$ as discussed in Proposition 1. In a more favorable setting, when the beliefs of the workers are not arbitrary but are aligned with the real observation probabilities and the mechanism has access to δ^*, it can be shown that for any $n \geq 2$, the platform makes strictly positive relative savings given by $P(1)\big(1 - \frac{1}{n}\big)$. We refer the reader to Theorem 2A of the appendix for the details.

We conclude the section with the following observation. While the employment of the PTSC mechanism with an appropriate scaling constant can guarantee that truth-telling is an equilibrium strategy, it is not hard to see that the denial strategy is still an equilibrium strategy in addition to

[6] In the original settings for which it was proposed, in which outcome-dependent lying incentives were not present (Radanovic, Faltings, and Jurca 2016).

the truth-telling strategy. Moreover, this undesired equilibrium is actually more profitable for the agents than the truth-telling equilibrium and any attempts of making the truth-telling equilibrium more profitable are impaired by the following result.

Proposition 2. *The denial strategy is an equilibrium of PTSC and is more profitable for the agents than the truth-telling equilibrium. More generally, if the denial strategy equilibrium exists in a mechanism, it is not possible to make the truth-telling equilibrium more profitable without causing loss to the platform.*

This type of uninformed equilibria are present throughout the related literature (e.g., see (Freeman, Lahaie, and Pennock 2017)) and in the next section, we will discuss how they can be eliminated under reasonable assumptions.

4 Honest Agents

In many real-life platforms with many participants, it is natural to assume that at least a few of them will behave honestly, regardless of the monetary incentives that the platform provides. This can be attributed to several reasons; for example, to rational choices that are not explicitly captured by the payments, e.g., an interest in the well-being of society or some intrinsic utility from "doing the right thing", or even to some form of bounded-rationality (Rubinstein 1998) or risk-aversion. We show that the the undesirable equilibrium highlighted in the previous section can be eliminated in our setting if it is known that there exists an arbitrary small non-zero fraction f of honest agents on the platform. In fact, it is only necessary that the agents *believe* that there is such a fraction of honest agents, which is a reasonable assumption in most real-world platforms. As it will be evident later, neither the rational agents nor the platform know the identity of the honest agents. Only assuming the existence of honest agents (without known identities) is fundamentally different from using identified trusted authorities for obtaining observations, since the latter violates the decentralization of the platform, while the former does not.

For the analysis, we will use an alternative definition of the self-predictor value that we defined in Section 3. This definition adapts the self-predictor value to the situation when agents believe that only a f-fraction of other agents are honest and the remaining $(1-f)$-fraction always report 0 irrespective of their observations, i.e. they follow the denial strategy.

Definition 5 (Self-Predictor Value With Colluding Agents)**.** *Let* $Q_i(X_p = 0|X_i = 1) = (1-f)+f\cdot P_i(X_p = 0|X_i = 1)$ *and* $Q_i(X_p = 0) = (1-f) + f \cdot P_i(X_p = 0)$. *The* self-predictor value with colluding agents *is defined as*

$$\delta_c^* = \min_i \left(\frac{P_i(X_p = 1|X_i = 1)}{P_i(X_p = 1)} - \frac{Q_i(X_p = 0|X_i = 1)}{Q_i(X_p = 0)} \right)$$

Note that when $f = 1$, we obtain exactly the same quantity as in Definition 4.

Lemma 1. *If* $\delta^* > 0$*, then* $\delta_c^* > 0$*, for any* $0 < f < 1$.

We will exploit this property of δ_c^* to show that it is possible to eliminate the denial strategy equilibrium for any non-zero value of f. Similar to the previous section, we assume that the mechanism knows only an estimate $\delta_c = \delta_c^* + \beta_c$.

Theorem 3. *Given that for* $f > 0$*, (a) an* f*-fraction of agents are honest, (b) the remaining* $(1-f)$*-fraction adopt the denial strategy and (c) it holds that* $\alpha > \frac{\mathcal{R}}{n\cdot\delta_c}$*, the truth-telling strategy is a strict best response if* $\beta_c \leq 0$ *and is an* $(\frac{\beta_c\cdot\mathcal{R}}{n\cdot\delta_c})$*-approximate best response if* $\beta_c > 0$.

The theorem implies that the collusion of the $(1-f)$-fraction who adopt the denial strategy becomes unstable and the rational choice for them will be to break the collusion and deviate to the truth-telling strategy. In other words, the denial strategy equilibrium is eliminated and the truthful equilibrium prevails.

Given that $\delta_c^* \leq \delta^*$ by definition (and strictly smaller when $f > 0$), the scaling constant α of PTSC in this case is actually larger than before. The reason is that we are now not only requiring that truth-telling is an equilibrium, but also that the denial strategy equilibrium is eliminated. Note that δ_c^* is strictly decreasing in f and achieves its maximum, which is δ^*, at $f = 1$.

For the saving, we first remark that the baseline for computing relative saving now naturally becomes the rational outcome in which the honest agents report the truth and the remaining agents play according to their denial strategies. Thus, the saving of a side-payment scheme, under which a total payment of $\mathcal{P}$ is made to *all* the agents (including the honest ones), now becomes:

$$\text{relative saving: } \frac{n\mathcal{R}' - \mathcal{P}}{n\mathcal{R}'},$$

where $\mathcal{R}' = \mathcal{R} \cdot \big[(1-f) + f \cdot (1 - P(1))\big]$. Note that $\big[(1-f) + f \cdot (1 - P(1))\big]$ is the expected value of the outcome when $(1-f)$-fraction of the agents play the denial strategy (always report 0) and the honest f-fraction report 0 only when they actually observe 0.

For the expected relative saving, we have the following.

Theorem 4. *If* $0 < f < 1$*, the expected relative saving made by the platform in the truth-telling equilibrium is at least*

$$\left[(1-f)P(1) - \frac{1}{n\delta_c}\right] \cdot \frac{1}{(1 - fP(1))}$$

Here, the lower bound on n needed for the saving to be positive is given by $n > \frac{1}{P(1)\cdot\delta_c\cdot(1-f)}$. Note that this lower bound depends inversely on $(1-f)$. If n is fixed, then one gets an upper bound on f given by

$$f < 1 - \frac{1}{P(1)\cdot\delta_c\cdot n}$$

An upper bound on f, or the direct dependence of n on f may seem counter-intuitive at first; why would one want to put a cap on the number of agents that always behave honestly? This is explained by the fact that these are merely the conditions required for a relative saving to be strictly positive. When there is a big enough fraction of honest agents,

the effect of the colluding agents on the outcome decreases and so does the relative saving that can be made by incentivizing these colluding agents to deviate to the truth-telling strategy. This means that if there are more honest agents than what the bound suggests (which tends to 1 for large n), then the platform will not actually save any money by implementing a side-payment mechanism. It should be noted however that Theorem 3 holds no matter how large f is, meaning that if the platform desires, at the expense of a negative saving, it can still implement the side-payment scheme in order to enforce that all agents are actually truth-telling in the equilibrium. The reason for wanting to do that could be to obtain correct information from the rational agents too, who would otherwise play denial strategy and introduce noise.

5 Experimental Evaluation

In this section, we evaluate the savings of PTSC experimentally on two real-world datasets, described below.

Dataset: We conducted experiments on the dataset[7] of (Zheng, Zhang, and Lyu 2010; 2014), which contains real-world Quality of Service evaluation results from 339 trusted agents on 5,825 web services. The agents observe the response time (in seconds) and throughput (in kbps) of the web-services and therefore, the observations can be used as two different datasets for our purposes. The dataset exhibits some missing observations but still has an overall density of 94.8% for the response time and 92.74% for the throughput. The observations are real values which we placed into two categories, corresponding to "good" and "bad" performance, in order to fit them to our binary observation setting. In particular, we treated a response time of at most 1 second as a "good" response time and the rest as "bad" response times. This resulted in 83.71% good response time observations, on average across all services. Similarly, we treated a throughput above 5 kbps as a good throughput and anything below that as a bad throughput. This resulted in 78.18% good throughput observations, on average across all services. Thus, in the context of our model, $P(1) \approx 0.8371$ for response time and $P(1) \approx 0.7818$ for throughput.

Simulation Parameters: We are interested in simulating settings in which the observations in the dataset would have been made by self-interested agents (rather than trusted ones) who have an incentive to play the denial strategy. Therefore, the dataset acts as the *true* private observations of the agents, which they may or may not reveal truthfully to the platform depending on their incentives. We fix a constant refund amount $\mathcal{R}$ in our simulations; since we will only discuss the relative saving, the actual choice of $\mathcal{R}$ is not important here. We vary the number of agents that are asked to report their observations for a service, by randomly selecting a subset of the agents from the dataset for every web-service.

We approximate the self-predictor value δ^* using the following process. We randomly sample, for each web service, two true observations. We use this sample to get an empirical estimate of the joint distribution of the observations of the agents and the prior distribution, and these two empirical estimates are used in the expression for δ^*. The result of this process can be thought of as a way to produce $\delta = \delta^* + \beta$, i.e., the value δ that appears in the statements of our theorems. As we mentioned in Section 3, since the value of δ^* is calculated as a minimum over all the agents, overestimating this value might cause some agents to have incentives to deviate, and in particular switch to their denial strategies. To examine the robustness of our scheme against this phenomenon, we quantify the savings of the mechanism when a fraction of agents, even with PTSC implemented, decide to play according to the denial strategy. We explain the results of these experiments in the next subsection.

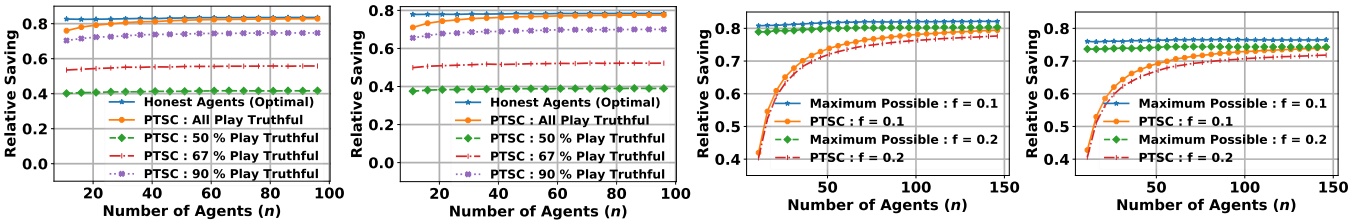
[7]Dataset is available at `http://wsdream.github.io`.

Experimental Results

In Figures 1a and 1b, we compare the saving achieved by PTSC against the optimal saving, which is obtained when all the agents are honest. Specifically, the optimal saving is given by $(\mathcal{P}_d - \mathcal{P}_\alpha)/\mathcal{P}_d$, whereas the saving of PTSC is given by $(\mathcal{P}_d - \mathcal{P}_{eq})/\mathcal{P}_d$, where $\mathcal{P}_d$ is the refund payment of the denial strategy equilibrium, $\mathcal{P}_\alpha$ is the refund payment when all agents are honest and $\mathcal{P}_{eq}$ is the total payment of PTSC, including the refund and side-payments. In line with our theoretical observation in Theorem 2, the saving achieved by PTSC converges the optimal saving, which is approximately $P(1)$, as the number of agents increases. In fact, the saving approaches the optimal levels quite quickly, for reasonable numbers of agents (i.e., approximately 40 agents). To quantify the robustness of PTSC with respect to the estimation of δ^*, the figure also depicts the relative saving made when only a 90%, 67% and 50% fraction of the agents receive the PTSC side-payments and report truthfully, and the rest receive the PTSC side-payment but still use the denial strategy. While the saving naturally declines, we observe that even with 90% of the agents being truthful, a significant relative saving is achieved.

We also consider the relative saving of PTSC when the side-payments are large enough to not only make truth-telling an equilibrium, but to also eliminate the denial strategy equilibrium, as discussed in Section 4, via the assumption that there exists an f-fraction of honest agents who always report truthfully. We set the value of f to either 0.1 or 0.2, and observe how quickly the relative savings made by PTSC can reach the maximum achievable relative saving as the number of agents increase; this is shown in Figures 1c and 1d. Note that unlike Figures 1a and 1b, here the relative saving starts at a lower value; this is because the scaling constant and hence the payment made by PTSC are required to be larger as discussed after Theorem 3. Also, note that we have a different maximum possible saving bound for each f. This is in agreement with our discussion at the end of Section 4 (following Theorem 4) i.e., a larger value of f decreases the maximum achievable relative saving.

While the relative saving only measures the monetary utility, the impact of PTSC on the correct outcome determination is also worth noting. Table 1 shows that as more agents are encouraged by the side payments to adopt the truth-telling strategy, the average outcome across all services approaches its true value, i.e., $1 - P(1)$.



(a) Response Time Data (b) Throughput Data (c) Response Time Data (d) Throughput Data

Figure 1: Relative saving made by PTSC.

Truthful agents %	Response Time	Throughput
50% truthful	0.5844	0.6093
67% truthful	0.4426	0.4765
90% truthful	0.2515	0.2964
All truthful	0.1684	0.2179

Table 1: Average value of the outcome as a function of the percentage of agents who are truth-telling, as a result of the side payments of PTSC. Note that "truthful" here does not refer to honest agents, but the agents that are incentivized by the mechanism to report their response times truthfully.

6 Conclusions and Future work

In this paper, we studied settings where the outcome is determined as an aggregate statistic of the reports of rational agents, who have outside incentives to manipulate the outcome in their favor.We discussed (i) how a detail-free peer-prediction mechanism, the PTSC mechanism, can be implemented as a side-payment scheme in order to guarantee that truth-telling is an equilibrium of the induced game and the undesired equilibrium, where all agents report a bad service, can be eliminated; and (ii) lower bounds on the relative saving in the net payments achieved by the mechanism, which approach optimality as the number of agents grows large.

As future work, it would be interesting to consider different outcome determination functions; threshold functions seem like an obvious choice, where the outcome is deemed "bad", when the fraction of agents that report that exceeds a given threshold. Given that PTSC provides guarantees for more general signal spaces, it would be interesting to study similar settings beyond the binary signal setting. However, that seems to require a different model of outcome determination and compensation schemes.

7 Appendix

Better Savings with Realistic Beliefs

Let us call the beliefs of the agents *realistic* if for any agent i, it holds that $P_i(X_p = x) = P(x)$ and $P_i(X_p = y|X_i = x) = P(y|x)$, where $P(x)$ is the real probability of an agent observing 0 for any random question and $P(y|x)$ is the real probability that for any given question, given that one agent reported x, the other reported y.

Theorem 2A. *Given that the beliefs of the agents are realistic and $\delta = \delta^*$, then the relative saving in the PTSC truthful equilibrium is given by:*

$$P(1)\left[1 - \frac{1}{n}\right]$$

Note that the relative saving in this case is strictly positive for any $n > 1$ and also approaches the optimal value of $P(1)$ as $n \to \infty$.

Covering the payments with the revenue from subscription fees

In the main text, we quantified the savings in relation to the amount that the platform would have to pay, if no side payment scheme was not in place and the agents would rationally follow their denial strategies. A slightly different way to quantify the saving is in terms of whether the revenue generated by the subscription fees is enough to cover the payments. In the web services example, the subscription fees is the amount of money that the agents pay to get the service (and the SLA) in the first place. Given refund $\mathcal{R}$, the *subscription fee* to the service will be $\mathcal{R} \cdot t$ for some $t \in \mathbb{R}$. Obviously, if we do not implement a side payment scheme and we let agents play their denial strategies, the payments can be covered with a $(1/t)$-fraction of the revenue, only if $t \geq 1$, and they can not be covered otherwise. On the other hand, the best that we can hope for is again given by the case when all the agents are honest. This is formalized below:

Proposition 3. *If agents were honest and $t \geq 1$, the refund payments can be covered in expectation with $\left(\frac{1-P(1)}{t}\right)$-fraction of the revenue generated. If $t < 1$, the refund payments can still be covered in expectation as long as $t \geq 1 - P(1)$.*

When the agents are rational and we implement the PTSC as side-payment mechanism, we have the following theorem.

Theorem 5. *In the truth-telling equilibrium of PTSC and assuming $n > \frac{1}{P(1)\cdot\delta}$, if $t \geq 1$ the payments can be covered in expectation with $\frac{1}{t}\left(1 - P(1) + \frac{1}{n\cdot\delta}\right)$-fraction of the revenue generated. If $t < 1$, the payments can still be covered in expectation, if $t \geq 1 - P(1) + \frac{1}{n\cdot\delta}$.*

As $n \to \infty$, the expressions for the fraction of revenue and minimum value t converges to the optimal values (Proposition 3).

References

Agarwal, A.; Mandal, D.; Parkes, D. C.; and Shah, N. 2017. Peer prediction with heterogeneous users. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC-2017)*.

Chakraborty, M., and Das, S. 2016. Trading on a rigged game: Outcome manipulation in prediction markets. In *IJCAI*, 158–164.

Chakraborty, M.; Das, S.; Lavoie, A.; Magdon-Ismail, M.; and Naamad, Y. 2013. Instructor rating markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 159–165. AAAI Press.

Chen, Y.; Gao, X. A.; Goldstein, R.; and Kash, I. A. 2011. Market manipulation with outside incentives. In *AAAI*.

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, 319–330. ACM.

De Alfaro, L.; Shavlovsky, M.; and Polychronopoulos, V. 2016. Incentives for truthful peer grading. *arXiv preprint arXiv:1604.03178*.

Faltings, B., and Radanovic, G. 2017. Game theory for data science: Eliciting truthful information. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11(2):1–151.

Freeman, R.; Lahaie, S.; and Pennock, D. M. 2017. Crowdsourced outcome determination in prediction markets. In *AAAI*, 523–529.

Gao, A.; Wright, J. R.; and Leyton-Brown, K. 2016. Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse. In *2nd Workshop on Algorithmic Game Theory and Data Science at EC 2016*.

Gnosis. 2017. Gnosis whitepaper. `https://gnosis.pm/assets/pdf/gnosis-whitepaper.pdf`. White paper.

Goel, N., and Faltings, B. 2018. Deep bayesian trust: A dominant and fair incentive mechanism for crowd. *To appaear in Proceedings of the AAAI Conference on Artificial Intelligence 2019. arXiv preprint arXiv:1804.05560*.

Jurca, R., and Faltings, B. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *International Workshop on Internet and Network Economics*, 268–277. Springer.

Jurca, R., and Faltings, B. 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce*, 190–199. ACM.

Jurca, R.; Faltings, B.; and Binder, W. 2007. Reliable qos monitoring based on client feedback. In *Proceedings of the 16th international conference on World Wide Web*, 1003–1012. ACM.

Kamble, V.; Shah, N.; Marn, D.; Parekh, A.; and Ramachandran, K. 2015. Truth serums for massively crowdsourced evaluation tasks. *arXiv preprint arXiv:1507.07045*.

Kandy, V., and Loomb, S. 2017. Decibel.live: A decentralized noise pollution monitoring and incentive platform. `https://www.decibel.live/`.

Kong, Y., and Schoenebeck, G. 2018. Equilibrium selection in information elicitation without verification via information monotonicity. *Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS)*.

Liu, Y., and Chen, Y. 2017a. Machine-learning aided peer prediction. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, 63–80. ACM.

Liu, Y., and Chen, Y. 2017b. Sequential peer prediction: Learning to elicit effort using posted prices. In *AAAI*, 607–613.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51(9):1359–1373.

Peck, M. E. 2017. Blockchains: How they work and why they'll change the world. *IEEE spectrum* 54(10):26–35.

Peterson, J.; Krug, J.; Zoltu, M.; Williams, A. K.; and Alexander, S. 2018. Augur: a decentralized oracle and prediction market platform. White paper.

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

Radanovic, G., and Faltings, B. 2013. A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI" 13)*, number EPFL-CONF-197486, 833–839.

Radanovic, G., and Faltings, B. 2015. Incentive schemes for participatory sensing. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1081–1089. International Foundation for Autonomous Agents and Multiagent Systems.

Radanovic, G.; Faltings, B.; and Jurca, R. 2016. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7(4):48.

Rubinstein, A. 1998. *Modeling bounded rationality*. MIT press.

Shnayder, V.; Agarwal, A.; Frongillo, R.; and Parkes, D. C. 2016. Informed truthfulness in multi-task peer prediction. EC '16, 179–196. ACM.

Waggoner, B., and Chen, Y. 2014. Output agreement mechanisms and common knowledge. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Witkowski, J., and Parkes, D. C. 2012a. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 964–981. ACM.

Witkowski, J., and Parkes, D. C. 2012b. A robust bayesian truth serum for small populations. In *AAAI*.

Zheng, Z.; Zhang, Y.; and Lyu, M. R. 2010. Distributed qos evaluation for real-world web services. In *2010 IEEE International Conference on Web Services*, 83–90. IEEE.

Zheng, Z.; Zhang, Y.; and Lyu, M. R. 2014. Investigating qos of real-world web services. *IEEE transactions on services computing* 7(1):32–39.