Peer-Prediction in the Presence of Outcome Dependent Lying Incentives

Naman Goel[1] , **Aris Filos-Ratsikas**[2] and **Boi Faltings**[1]
[1]Swiss Federal Institute of Technology, Lausanne (EPFL)
[2]University of Liverpool, UK
{naman.goel, boi.faltings}@epfl.ch, aris.filos-ratsikas@liverpool.ac.uk

Abstract

We derive conditions under which a peer-consistency 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 public real datasets.

1 Introduction

The task of eliciting reliable information from the participants of a system or a platform is rather fundamental and quite challenging: How can we incentivize the participating agents to provide the information they observe, *truthfully*? For example, imagine that we are trying to determine the quality of a service (e.g., a streaming service or an internet service) based on the feedback provided by the users. How can we trust that the users will spend the effort to come up with informative feedback, that actually reflect the quality of the service, rather than random numbers?

With access to a ground truth, i.e., a known measure of quality, a set of *proper scoring rules* are known to be capable of inducing truthful behavior from the users. But what about settings like the one of the example above, where we do not have an objective measure of the quality of the service, or that information is too costly to obtain? *Peer-prediction* [Miller *et al.*, 2005; Prelec, 2004] (and in general, *peer consistency* [Faltings and Radanovic, 2017]) mechanisms deal with precisely this issue: the broad idea is to match the report of an agent with that of a randomly chosen peer, and provide a payment as a function of the two reports.

The literature on peer-consistency [Faltings and Radanovic, 2017]) is quite rich. This literature includes solutions that are guaranteed to incentivize truth-telling even when there is a *cost of effort* for forming an informed report [Radanovic *et al.*, 2016], but it does not address settings in which agents have other incentives dependent on the aggregate feedback. Such cases arise frequently, for example, in decentralized QoS monitoring, environmental data collection and surveys that inform policy-making.

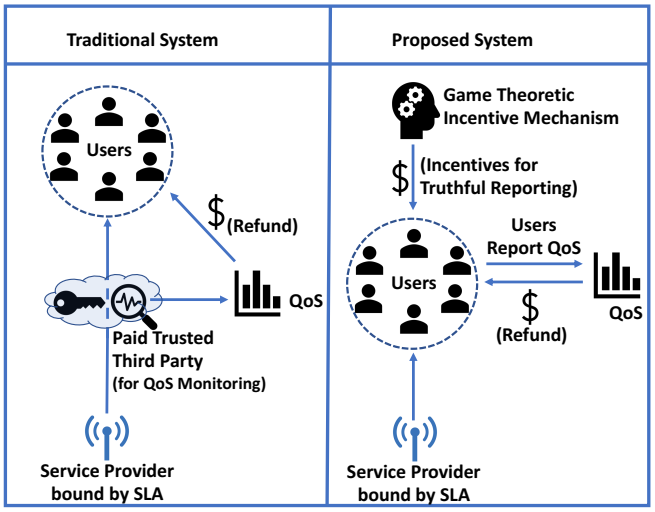


Figure 1: QoS Monitoring: A Motivating Example of Outcome Dependent Lying Incentives. Replacing trusted third party with the users themselves creates lying incentives for the users to misreport the true QoS received from the service provider. A game-theoretic mechanism incentivizes the users to report truthfully. As shown by [Goel *et al.*, 2020], the game-theoretic mechanism itself can be implemented by a regulatory authority in a completely decentralized, trustless and transparent manner by using the blockchain technology. A smart contract can automatically process refunds based on the crowdsourced outcome and the conditions of the SLA.

As a concrete 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, where the agreement dictates that the client will be compensated if there is a violation. Traditionally, a trusted third party monitors the quality of service (QoS) and sends the reports to a regulatory authority (or the service provider itself for self-regulation). Depending on the collected reports from the trusted party and the conditions of the SLA, users are issued refund. Not only this traditional approach is costly due to the high cost of hiring a commercial third party but it is also not a transparent

and decentralized approach from the users' perspective. On the other hand, if we were to decide whether such a violation actually took place based on the feedback from the clients themselves, it is clear that the clients would have incentives to report a violation regardless, in order to be compensated. To get truthful reports from the users, their incentives must be aligned with truthful behavior by using a game-theoretic mechanism. The main question here is, can we still use a peer-consistency mechanisms to counter-act this type of *outside* incentives?

1.1 Our Contributions

In this paper, we consider settings with outside incentives and binary observations, and we employ the PTSC mechanism of [Radanovic *et al.*, 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. This is the first result of this nature, that shows that a peer-consistency mechanism can be applied to the case of outcome-dependent incentives in a general information elicitation scenario.

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. 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 *et al.*, 2016]; our paper extends the analysis of PTSC in the *presence* of the outside 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 one has to not only ensure a truthful equilibrium, but also 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-consistency (e.g. see [Jurca and Faltings, 2005], [Witkowski and Parkes, 2012]) and translates to positive correlation 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.

1.2 Related Work

Our work draws on the recent ideas in the peer-consistency literature [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 *et al.*, 2016; Gao *et al.*, 2016; Agarwal *et al.*, 2017; Liu and Chen, 2017; Kong and Schoenebeck, 2018; Goel and Faltings, 2019a; Goel and Faltings, 2019b]; 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 behavior, 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 *et al.*, 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-consistency 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 *et al.*, 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 *et al.*, 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 *et al.*, 2017] do not address the issue of non-truthful equilibria in their work. In settings other than prediction markets, [Jurca *et al.*, 2007] consider QoS feedback elicitation but assume full knowledge of agents' beliefs and only constant lying incentives.

Finally, we remark that, as we mentioned earlier, [Radanovic *et al.*, 2016] have already shown that the PTSC mechanism can be tuned to overcome the cost of effort that the agents might have for coming up with their observations, which is a constant quantity. This is therefore markedly different from the case of outside incentives, where the "cost" that needs to be overcome crucially depends on the reports of the other agents.

2 Model and Objectives

We consider settings in which questions are to be resolved on a decentralized platform through acquiring feedback from agents. No agent, however, is required to answer more than one 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 a question and submits her feedback report $Y_i \in \{0, 1\}$ to the platform. For any 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 not 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. We remark here that more generally, the outcome can be defined as any non-negative, non-decreasing function of the fraction of agents who report 0 (e.g., a *threshold* function that becomes 1 if, say, 70% of the agents report 0). We choose the fraction of dissatisfied agents as our outcome function, for the reasons mentioned above, and also following the related literature [Freeman *et al.*, 2017].

The main novelty of our setting is that the agents receive an outside incentive that is dependent on this aggregate outcome. More precisely, the payment given to an agent is $\mathcal{R} \cdot o_w$, where where $\mathcal{R}$ is a positive constant. In the web service example, such payments might arise through the service level agreements between the web service provider and the agents. The focus of this paper is how to adapt the incentives given for the reports to overcome such lying incentives.

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 may not report their true observations, if not properly incentivized to do so. We follow the common assumptions that agents are *risk-neutral* and that the utilities are *non-transferable*.

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 report. Some examples are as follows:s

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.

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 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 any 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, if such an outcome is used to issue the payments to the agents, it can incur a huge loss on the platform.

Peer-consistency. 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-consistency mechanisms* [Faltings and Radanovic, 2017] to align the incentives of the agents with their feedback. The most important constituents of the peer-consistency 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.[1] 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 $P_i(X_p = x'|X_i = x)$.

[1] 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.

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-consistency 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.

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 fares well against this outcome, will fare well against any other side-payment scheme, including one in which the outcome determination is done by a costly third party.

The PTSC Mechanism. Since we are interested in relaxing the informational assumptions as much as possible, we will use a detail-free mechanism (that doesn't know agents' beliefs) for determining the side-payments on the decentralized platform. Note that it is not necessary that a given user may be able to answer multiple questions (about different web services). This rules out several multi-task mechanisms like [Dasgupta and Ghosh, 2013; Shnayder *et al.*, 2016]. Thus, we will use the PTSC mechanism [Radanovic *et al.*, 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:

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

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. The PTSC mechanism is a special case of the PTS mechanism [Jurca and Faltings, 2011] and is based on the idea of using $R_i(y)$ from other apriori similar questions to estimate the prior belief of the users. It is possible to use other ways to estimate the prior in the PTS mechanism and relax the requirement of having other 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-consistency 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 *$\varepsilon$-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 any other strategy σ_i' by at most ε. Since beliefs need not be common among agents, i.e. they are subjective, the equilibrium concept that we adopt is the *ex-post subjective equilibrium* [Witkowski and Parkes, 2012]. In this equilibrium concept, an agent's best response is independent of the beliefs of others. In the paper, we will 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-consistency for binary answer spaces, e.g. see [Jurca and Faltings, 2005; Witkowski and Parkes, 2012] and it is under this condition that PTSC guarantees that truth-telling is an equilibrium.[2] We will make the same assumption throughout this paper, and we

[2] In the original settings for which it was proposed [Radanovic *et al.*, 2016], outcome-dependent lying incentives were not present.

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 *et al.*, 2016] capturing similar 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* in [Freeman *et al.*, 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.

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.*

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 $P(1)$ 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 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 if no further assumption are made about agents' beliefs, the denial strategy is still an equilibrium strategy in addition to the truth-telling strategy. [Radanovic *et al.*, 2016] have shown that when outcome dependent lying incentives are not present, while this uninformed equilibria does exist in PTSC, it is not profitable (pays zero expected reward). Unfortunately, in the presence of outcome dependent lying incentives, this undesired equilibrium becomes more profitable than the truth-telling equilibrium because every agent can now get the maximum value of the refund $\mathcal{R}$ by playing the denial strategy. Any attempts of making the truth-telling equilibrium more profitable in this setting are impaired by the following result.

Proposition 2. *If the denial strategy equilibrium exists in any mechanism in the presence of outcome dependent lying incentives, it is not possible to make the truth-telling equilibrium more profitable without causing loss to the platform.*

Here loss means that the total payment will be higher than the maximum refund $\mathcal{R}$. While this negative result is reminiscent of the known negative result about uninformative equilibria in the peer prediction mechanism of [Miller *et al.*, 2005] reported in [Jurca and Faltings, 2005; Jurca and Faltings, 2009], in our result focal uninformative equilibria arise because of outside incentives and not due to a weakness of the incentive mechanism.

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 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 (as proposed in [Jurca and Faltings, 2005]), 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 equilibrium is eliminated and the truthful equilibrium prevails. Thus, we get the following proposition.

Proposition 3. *Under the conditions derived in Theorem 3, the denial strategy is no longer an equilibrium strategy.*

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 benchmark against which we compare now naturally becomes the rational outcome in which the honest agents report the truth and the remaining agents play according to their denial strategies. Concretely, the saving of a side-payment scheme, under which a total payment of $\mathcal{P}$ are made to the agents, 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.

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

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

We 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 and the above theorem has been derived accounting for this fact. In theorem 4, 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. It is further shown in the proof that the relative saving in this case too approaches the optimal relative saving as $n \to \infty$.

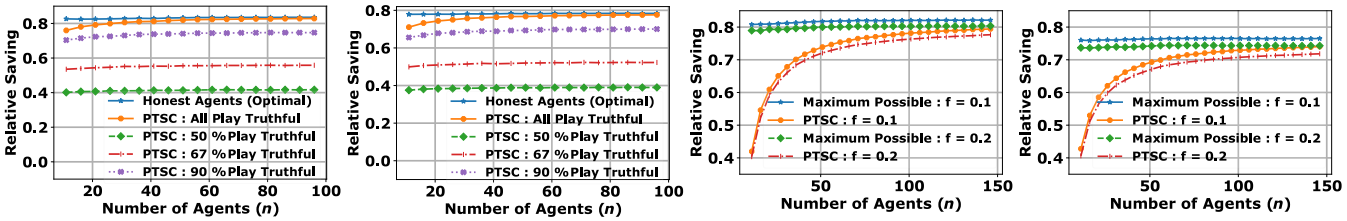
5 Experimental Evaluation

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

Dataset. We conducted experiments on the dataset[3] of [Zheng *et al.*, 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 response time and 92.74% for 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. We treated a response time of at most 1 second as a "good" response time and the rest as "bad". 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. 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.

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



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

Figure 2: Relative saving made by PTSC.

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, play the denial strategy.

5.1 Experimental Results

In Figures 2a and 2b, 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, we achieve a significant saving.

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, assuming 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 2c and 2d. Note that unlike Figures 2a and 2b, 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 following Theorem 4 i.e., larger values of f lower the maximum achievable relative saving.

6 Conclusions

In this paper, we studied settings motivated by polls and other crowdsourced data where the agents reporting the data have a conflict of interest with the aggregate statistic of the reported data. Such scenarios occur for example in reporting environmental data, where some reports might downplay pollution, in polls such as the LIBOR, where reporting agents have direct financial interest in the poll result, or our running example of self-reported Quality of Service measurements, where reporters may hope for refunds. We showed

(i) how a detail-free peer-consistency mechanism, the PTSC mechanism, can be implemented to guarantee that truth-telling is an equilibrium of the induced game, in spite of the outside incentives to the contrary,

(ii) how the presence of honest agents, which can remain anonymous, eliminates the undesired equilibrium where all agents report the outcome that benefits their outside incentive; and

(iii) lower bounds on the relative saving in the net payments achieved by the mechanism, which approach optimality as the number of agents grows large.

We only considered a scenario where the outside incentives favor the same misreport for all agents, and do so with a particular dependence on the outcome. In ongoing work, we are considering different forms of outcome dependence, in particular threshold functions that require that the outcome exceeds a given threshold for the users to get refunds, and it turns out that these lead to different results. In the future, it would also be interesting to consider cases where agents have different and possibly opposing interests, such as in polls where different populations want different outcomes to win. Given that PTSC provides guarantees for non-binary signal spaces too, it would also be interesting to study similar problem beyond the binary answer setting. However, that seems to require somewhat different formalization for the correct determination of the outcome and the compensation schemes.

References

[Agarwal *et al.*, 2017] Arpit Agarwal, Debmalya Mandal, David C. Parkes, and Nisarg Shah. Peer prediction with heterogeneous users. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)*, 2017.

[Chakraborty and Das, 2016] Mithun Chakraborty and Sanmay Das. Trading on a rigged game: Outcome manipulation in prediction markets. In *IJCAI*, 2016.

[Chakraborty *et al.*, 2013] Mithun Chakraborty, Sanmay Das, Allen Lavoie, Malik Magdon-Ismail, and Yonatan Naamad. Instructor rating markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, 2013.

[Chen *et al.*, 2011] Yiling Chen, Xi Alice Gao, Rick Goldstein, and Ian A Kash. Market manipulation with outside incentives. In *AAAI*, 2011.

[Dasgupta and Ghosh, 2013] Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.

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

[Freeman *et al.*, 2017] Rupert Freeman, Sébastien Lahaie, and David M Pennock. Crowdsourced outcome determination in prediction markets. In *AAAI*, 2017.

[Gao *et al.*, 2016] Alice Gao, James R Wright, and Kevin Leyton-Brown. 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.*, 2016.

[Goel and Faltings, 2019a] Naman Goel and Boi Faltings. Deep bayesian trust: A dominant and fair incentive mechanism for crowd. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[Goel and Faltings, 2019b] Naman Goel and Boi Faltings. Personalized peer truth serum for eliciting multi-attribute personal data. In *35th Conference on Uncertainty in Artificial Intelligence, Tel Aviv (UAI 2019)*. AUAI, 2019.

[Goel *et al.*, 2020] Naman Goel, Cyril van Schreven, Aris Filos-Ratsikas, and Boi Faltings. Infochain: A decentralized, trustless and transparent oracle on blockchain. *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[Jurca and Faltings, 2005] Radu Jurca and Boi Faltings. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *International Workshop on Internet and Network Economics*, pages 268–277. Springer, 2005.

[Jurca and Faltings, 2009] Radu Jurca and Boi Faltings. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34:209–253, 2009.

[Jurca and Faltings, 2011] Radu Jurca and Boi Faltings. Incentives for answering hypothetical questions. In *The 1st Workshop on Social Computing and User Generated Content, EC*, 2011.

[Jurca *et al.*, 2007] Radu Jurca, Boi Faltings, and Walter Binder. Reliable qos monitoring based on client feedback. In *Proceedings of the 16th international conference on World Wide Web*, pages 1003–1012. ACM, 2007.

[Kamble *et al.*, 2015] Vijay Kamble, Nihar Shah, David Marn, Abhay Parekh, and Kannan Ramachandran. Truth serums for massively crowdsourced evaluation tasks. *arXiv preprint arXiv:1507.07045*, 2015.

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

[Liu and Chen, 2017] Yang Liu and Yiling Chen. Machine-learning aided peer prediction. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 63–80. ACM, 2017.

[Miller *et al.*, 2005] Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 2005.

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

[Radanovic and Faltings, 2013] Goran Radanovic and Boi Faltings. A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI" 13)*, 2013.

[Radanovic and Faltings, 2015] Goran Radanovic and Boi Faltings. Incentive schemes for participatory sensing. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1081–1089. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[Radanovic *et al.*, 2016] Goran Radanovic, Boi Faltings, and Radu Jurca. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):48, 2016.

[Rubinstein, 1998] Ariel Rubinstein. *Modeling bounded rationality*. MIT press, 1998.

[Shnayder *et al.*, 2016] Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. Informed truthfulness in multi-task peer prediction. EC '16. ACM, 2016.

[Waggoner and Chen, 2014] Bo Waggoner and Yiling Chen. Output agreement mechanisms and common knowledge. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

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

[Zheng *et al.*, 2014] Zibin Zheng, Yilei Zhang, and Michael R Lyu. Investigating qos of real-world web services. *IEEE transactions on services computing*, 2014.