Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Incentives for Subjective Evaluations with Private Beliefs

Goran Radanovic and **Boi Faltings**
Ecole Polytechnique Federale de Lausanne
(EPFL)
Artificial Intelligence Laboratory
CH-1015 Lausanne, Switzerland
{goran.radanovic, boi.faltings}@epfl.ch

Abstract

The modern web critically depends on aggregation of information from self-interested agents, for example opinion polls, product ratings, or crowdsourcing. We consider a setting where multiple objects (questions, products, tasks) are evaluated by a group of agents. We first construct a minimal peer prediction mechanism that elicits honest evaluations from a homogeneous population of agents with different private beliefs. Second, we show that it is impossible to strictly elicit honest evaluations from a heterogeneous group of agents with different private beliefs. Nevertheless, we provide a modified version of a divergence-based Bayesian Truth Serum that incentivizes agents to report consistently, making truthful reporting a weak equilibrium of the mechanism.

Introduction

The main idea behind a participatory web approach is that any participant can provide her private information which is then aggregated into the web content. There are numerous examples, ranging from hotel reviews on TripAdvisor[1] to human intelligence tasks on MTurk[2]. Since private information cannot be easily verified, one of the key challenges is to incentivize participants to invest effort and provide their true opinions.

We model this scenario with a group of agents evaluating a set of similar objects. The agents represent rational participants, while objects can be anything that the participants can provide their opinions about: questions in case of opinion polling, product evaluations in case of product reviewing, or subjective tasks in case of crowdsourcing.

The standard information elicitation techniques often score responses against the ground truth, as it is for proper scoring rules or prediction markets (Savage 1971; Gneiting and Raftery 2007; Lambert and Shoham 2009; Hanson 2003; Chen and Pennock 2007). However, in our scenario the ground truth is not known to a mechanism designer, or might not even be well defined.

[1] http://www.tripadvisor.com

[2] https://www.mturk.com/

Instead of using the ground truth, peer evaluation techniques score an agent against her *peers*, i.e. against agents who also participate in the evaluation process. Many peer evaluation methods have been developed in recent years, most of them fitting one of the following categories.

Knowledge dependent mechanisms require knowledge about agents' beliefs. The peer prediction methods (Miller, Resnick, and Zeckhauser 2005; Jurca and Faltings 2006; 2007) are representatives of this category. These mechanisms are minimal in a sense that they only elicit desired information. The collective revelation of (Goel, Reeves, and Pennock 2009) can also be placed in this category.

Common prior mechanisms are a group of mechanisms that assume common prior belief among agents, but this prior does not need to be known to the mechanism. A typical representative of this group is the Bayesian Truth Serum (Prelec 2004), with its variants (Witkowski and Parkes 2012b; Radanovic and Faltings 2013; 2014), in which agents provide an additional report along with the information that the mechanism wants to elicit. In this group, we can also include the knowledge-free mechanism of (Zhang and Chen 2014) and the minimum truth serum of (Riley 2014).

The *private prior* mechanism of (Witkowski and Parkes 2012a) for elicitation of binary information does not require knowledge about agents' beliefs nor do the agents have to have a common prior. However, the mechanism assumes a temporal separation in the elicitation process, i.e. agents first provide their prior beliefs to the mechanism, and then they make evaluations that they report to the mechanism.

Weakly truthful mechanisms do not necessarily provide strong incentives for truthfulness (Lambert and Shoham 2008) or may not necessarily be truthful (Jurca and Faltings 2011). In this group of mechanisms, we can also add the output agreement mechanism of (Waggoner and Chen 2013; 2014) that elicits common knowledge, rather than agents' private information.

It is not surprising that all the aforementioned mechanisms require a certain restriction on either: knowledge about agents' beliefs, homogeneity of agents' beliefs, structure of elicitation process or structure of the information being elicited. In fact, several impossibility results (Jurca and Faltings 2011; Waggoner and Chen 2013; 2014; Radanovic and Faltings 2013; 2014) indicate that in a single shot elicitation process, one cannot do much better.

It is often the case, however, that one wants to elicit information about several statistically similar objects from the same group of agents. For example, in product reviewing, a certain group of agents might rate several products that are a priori similar to each other. Another example is crowdsourcing where a single agent solves multiple (subjective) tasks.

These type of settings have already been analyzed by (Witkowski and Parkes 2013) and (Dasgupta and Ghosh 2013) that introduce private prior mechanisms for elicitation of binary information. The key difference between the two settings is that (Witkowski and Parkes 2013) assumes that agents receive their private information in a similar fashion, while (Dasgupta and Ghosh 2013) does not. In other words, for the former setting, agents have homogeneous characteristics, although their beliefs are private, while for the latter, agents are considered to be entirely heterogeneous, both in how they obtain their private information and in their beliefs.

We build on their ideas and construct two private prior mechanisms, one for a homogeneous population of agents, and another for a heterogeneous population. For a homogeneous population, we show a minimal mechanism that requires only a single report. Unlike the mechanism from (Witkowski and Parkes 2013), our minimal mechanism does not require a knowledge of a *belief change bound*, and can be applied to non-binary domains.

For heterogeneous populations, we show that for non-binary values, no minimal mechanism is incentive compatible in general case. The mechanism in (Dasgupta and Ghosh 2013) can achieve truthfulness because it is restricted to elicitation of binary information, and makes additional assumptions. We show another alternative that works for any number of values but is not minimal, i.e. requires an additional prediction report. It modifies the approach of (Radanovic and Faltings 2014) to encourage *consistent* (non-random) reports. Since truthful reporting is a consistent behaviour, the novel mechanism is (weakly) incentive compatible.

The Setting

We are interested in a scenario where a group of agents are asked to provide their opinions regarding a priori similar objects. We follow the usual peer prediction setting, where agents' opinions are formed stochastically. Our setting has the following structure.

There are $M >> 1$ different objects $\{o_1, ..., o_M\}$ with properties defined by a random variable Ω that takes values in a finite discrete set $\{\omega, ...\}$. For each object, Ω is generated stochastically according to a distribution $Q(\Omega = \omega)$.

When an agent evaluates an object o, she receives a private signal X^o that represents her evaluation of object o. The private signal takes values in $\{x, y, z, ...\}$ and is obtained by sampling a distribution $Q(X^o|\Omega = \omega)$. Naturally, this means that objects with the same properties Ω should generate the same histogram of opinions, though a single agent might have different opinion about different objects.

As in the standard peer prediction settings (e.g. (Miller, Resnick, and Zeckhauser 2005; Prelec 2004)), private signals $X^o_{a_1}$ and $X^o_{a_2}$ of two different agents a_1 and a_2 who evaluate the same object o are conditionally independent given $\Omega = \omega$, i.e. $Q(X^o_{a_2}|X^o_{a_1}, \Omega = \omega) = Q(X^o_{a_2}|\Omega = \omega)$. Moreover, distributions $Q(\Omega)$ and $Q(X^o|\Omega)$ are fully mixed, i.e. $Q(\Omega = \omega) > 0$ and $Q(X^o = x|\Omega = \omega) > 0$, while signals $X^o_{a_1}$ and $X^o_{a_2}$ are stochastically relevant, i.e. there exists z such that $Q(X^o_{a_2} = z|X^o_{a_1} = x) \neq Q(X^o_{a_2} = z|X^o_{a_1} = y)$ for $x \neq y$.

A mechanism does not have access to the true values of $Q(\Omega)$ and $Q(X|\Omega)$. On the other hand, agents form beliefs regarding the true values, and these beliefs can be different for different agents, i.e. agents do not have a common belief.

We distinguish two types of populations. In a *homogeneous* population, agents might have different private beliefs, but their private signals are generated in a similar fashion, i.e. $Q(X^o_{a_1}|\Omega = \omega) = Q(X^o_{a_2}|\Omega = \omega)$ for two different agents a_1 and a_2 who evaluate an object o of type ω. Again, notice that agents a_1 and a_2 might have different beliefs about distribution $Q(X|\Omega = \omega)$. In a *heterogeneous* population, agents might have different beliefs and their private signals might be generated using different probability distribution functions, i.e. $Q(X^o_{a_1}|\Omega = \omega)$ can differ from $Q(X^o_{a_2}|\Omega = \omega)$.

We denote an agent a's belief about true distributions $Q(X^o|\Omega)$ and $Q(\Omega)$ with $R_a(X^o|\Omega)$ and $R_a(\Omega)$. An agent a_1, who evaluated object o_1, has two important components in her belief system. The *prior* belief $Pr(X^{o_2}_{a_2} = x)$, also denoted by $Pr(x)$, is a belief regarding the private signal of another agent a_2 who evaluated object o_2 that is not evaluated by agent a_1. This belief can be easily calculated from $R_{a_1}(\Omega)$ and $R_{a_1}(X^o|\Omega)$ using:

$$Pr(X^{o_2}_{a_2} = x) = \sum_{\omega} R_{a_1}(X^{o_2}_{a_2} = x|\Omega = \omega) R_{a_1}(\Omega = \omega)$$

The *posterior* belief $Pr(X^{o_1}_{a_2} = y|X_{a_1} = x)$, also denoted by $Pr(y|x)$, is a belief regarding a private signal of another agent a_2, often called *peer* agent, that evaluated the same object as agent a_1. This belief can be calculated using the conditional independence of $X^{o_1}_{a_1}$ and $X^{o_1}_{a_2}$:

$$Pr(X^{o_1}_{a_2} = y|X^{o_1}_{a_1} = x) =$$

$$\sum_{\omega} R_{a_1}(X^{o_1}_{a_2} = y|X^{o_1}_{a_1} = x, \Omega = \omega) R_{a_1}(\Omega = \omega|X^{o_1}_{a_1} = x)$$

where by Bayes' rule:

$$R_{a_1}(\Omega|X^{o_1}_{a_1} = x) = \frac{R_{a_1}(X^{o_1}_{a_1} = x|\Omega) R_{a_1}(\Omega)}{Pr(X^{o_1}_{a_1} = x)}$$

$Pr(X^{o_1}_{a_1} = x)$ is agent a_1's prior belief regarding her own evaluation. In our analysis, Pr refers to a belief of an agent that is being scored, so we do not additionally index it.

Once an agent evaluates an object, she reports her private signal (evaluation) to the mechanism. Since an agent might lie, we denote her report by Y. Moreover, an agent might also be asked to report her prediction regarding the frequency of reports for an object she evaluated; we denote this prediction by $\mathbf{F}$. When an agent is honest, $\mathbf{F}$ corresponds to her posterior belief.

Quadratic Scoring Rule

One of the main tools for elicitation of private beliefs is a class of mechanisms called strictly proper scoring rules.

Suppose that an agent is asked to report her prediction (belief) $\mathbf{F}$ regarding an event whose outcome x eventually becomes known to the mechanism. For example, in a weather forecast, prediction $\mathbf{F}$ is a probability distribution function over possible weather conditions, while x is the realized weather condition. If the agent is rewarded with a strictly proper scoring rule $S(\mathbf{F}, x)$, her reward is in expectation maximized for reporting the true prediction.

In peer prediction settings, prediction $\mathbf{F}$ corresponds to an agent's posterior or prior belief Pr regarding private signals of the other agents. Notice that a proper scoring rule takes as an input an agent's belief (probability distribution function), and not her private signal (scalar value), so it cannot be directly applied in the elicitation of agents' private signals. However, agents' beliefs depend on their private signals, which can be utilized in the elicitation process.

There is a wide variety of strictly proper scoring rules, e.g. logarithmic, spherical or quadratic scoring rules (Gneiting and Raftery 2007). In this paper, we use the quadratic scoring rule defined as:

$$S(\mathbf{F}, x) = \frac{1}{2} + \mathbf{F}(x) - \frac{1}{2}\sum_{y} \mathbf{F}(y)^2 \qquad (1)$$

which produces bounded payoffs that take values in $[0, 1]$.

Homogeneous Populations

We first consider a homogeneous population of agents where opinions of different agents are formed in a similar fashion. Unlike (Witkowski and Parkes 2013), we do not try to learn (estimate) agents' priors, but rather construct an incentive scheme from finite and small number of samples.

We demonstrate that the elicitation of private signals can be done with a minimal mechanism that asks agents to report only their evaluations. We define proxy events linked to the evaluations of other agents whose probabilities are the same as those that would be reported in a prediction report, and use these to construct an expression that an agent expects to be the same as the quadratic scoring rule.

Minimal Peer Prediction with Private Priors

Consider an agent a_1 that is asked to provide her opinion regarding an object o_1. Once agent a_1 evaluates her object to $X^{o_1}_{a_1}$, she updates her belief regarding her *peer* agent a_2, who has also evaluated object o_1, to $Pr(\cdot|X^{o_1}_{a_1})$, where $\cdot$ denotes any possible evaluation $\{x, y, z, ...\}$. If agent a_1 believes that other agents are honest, $Pr(\cdot|X^{o_1}_{a_1})$ is also her belief about her peer's report.

Now, consider another object $o_2 \neq o_1$ evaluated by a third agent a_3 and not evaluated by agent a_1. Agent a_1's belief about the evaluation of agent a_3 is $Pr(\cdot)$. However, agents obtain their private signals in a similar fashion. So, if agent a_1 knows about a proxy agent a_{proxy} who evaluated object o_2 with y, agent a_1's belief about agent a_3's evaluation changes from $Pr(\cdot)$ to $Pr(\cdot|y)$. This means that the indicator variable $\mathbb{1}_{X^{o_2}_{a_3}=z}$ is in expectation equal to $Pr(z|y)$.

Now, suppose that a_3 is honest, i.e. $Y^{o_2}_{a_3} = X^{o_2}_{a_3}$, and that the evaluation of honest a_{proxy} is equal to agent a_1's report, i.e. $Y^{o_2}_{a_{proxy}} = X^{o_2}_{a_{proxy}} = Y^{o_1}_{a_1}$. Then, the indicator variable $\mathbb{1}_{Y^{o_2}_{a_3}=z}$ is in expectation equal to agent a_1's belief $Pr(z|Y^{o_1}_{a_1})$ that would make up her prediction report regarding her peer's evaluation.

The idea is to arrange indicators $\mathbb{1}_{Y^{o_2}_{a_3}=z}$ so that they correspond to the scoring rule (1), where prediction $Pr(\cdot|Y^{o_1}_{a_1})$ is scored by how well it predicts the report of peer a_2. Provided that the peer is honest, the expected score is maximized when prediction is equal to $Pr(\cdot|X^{o_1}_{a_1})$, which implies truthfulness of agent a_1.

Minimal Peer Prediction with Private Priors. The mechanism has the following structure:

1. Ask an agent a_1 and her peer a_2 to evaluate object o_1, and report their evaluations $Y^{o_1}_{a_1}$ and $Y^{o_1}_{a_2}$ to the mechanism.
2. Randomly sample one response for all objects $o \neq o_1$ that are not evaluated by agent a_1. We denote this sample by Σ and we call it *double-mixed* if it contains all possible values from $\{x, y, z, ...\}$ at least twice.
3. If sample Σ is not double-mixed, agent a_1's score is equal to $u(Y^{o_1}_{a_1}, Y^{o_1}_{a_2}) = 0$.
4. Otherwise, take two different objects $o_2 \neq o_3 \neq o_1$ whose Σ samples are equal to $Y^{o_1}_{a_1}$, and randomly select another sample for each of them to obtain two responses $Y^{o_2}_{a_3}$ and $Y^{o_3}_{a_4}$. Finally, the score of an agent a is equal to:

$$u(Y^{o_1}_{a_1}, Y^{o_1}_{a_2}) = \frac{1}{2} + \mathbb{1}_{Y^{o_2}_{a_3}=Y^{o_1}_{a_2}} - \frac{1}{2}\sum_{z} \mathbb{1}_{Y^{o_2}_{a_3}=z}\mathbb{1}_{Y^{o_3}_{a_4}=z}$$

Notice that the fourth step of the mechanism is only applied when Σ is double-mixed, and this is important to prevent potential bias towards more likely evaluations. Namely, if the fourth step is executed whenever Σ contains two reports equal to report $Y^{o_1}_{a_1}$, which is sufficient for the score $u(Y^{o_1}_{a_1}, Y^{o_1}_{a_2})$, agent a_1 might report dishonestly in the hope of increasing the probability of getting non-zero payoff.

To illustrate the principle of the minimal peer prediction, consider an agent a_1 with evaluation for object o_1 equal to $X^{o_1}_{a_1}$. Her belief about the report of her peer is $Pr(\cdot|X^{o_1}_{a_1})$, while her belief about the report of an agent who evaluated a different object is $Pr(\cdot)$. Here, $\cdot$ denotes any possible evaluation $\{x, y, z, ...\}$. The mechanism works as follows. If Σ is not double-mixed - which happens with probability strictly less than 1 for $|\Sigma| \geq 2|\{x, y, z, ...\}|$ - agent a_1's reward is 0. Otherwise, the mechanism searches in Σ for two objects o_2 and o_3 whose Σ samples are equal to $Y^{o_1}_{a_1}$. Since agent a_1 knows that samples of o_2 and o_3 are equal to $Y^{o_1}_{a_1}$, and agents observe signals in a similar way, agent a_1 updates her belief about the other evaluations of objects o_2 and o_3: $Pr(\cdot) \rightarrow Pr(\cdot|Y^{o_1}_{a_1})$. This means that a_1's belief about reports of agent a_3 (who evaluates o_2 and whose report is not in Σ) and agent a_4 (who evaluates o_3 and whose report is not in Σ) is equal to $Pr(\cdot|Y^{o_1}_{a_1})$.

Furthermore, the indicators $\mathbb{1}_{a_3=z}$ and $\mathbb{1}_{a_4=z}$ in score $u(Y^{o_1}_{a_1}, Y^{o_1}_{a_2})$ are in expectation equal to $Pr(z|Y^{o_1}_{a_1})$. Therefore, assuming that agents other than a_1 are honest and that sample Σ is double-mixed, the score is in expectation equivalent to the quadratic scoring rule $S(Pr(X^{o_1}_{a_2}|Y^{o_1}_{a_1}), X^{o_1}_{a_2})$, which is in expectation maximized for $Y^{o_1}_{a_1} = X^{o_1}_{a_1}$.

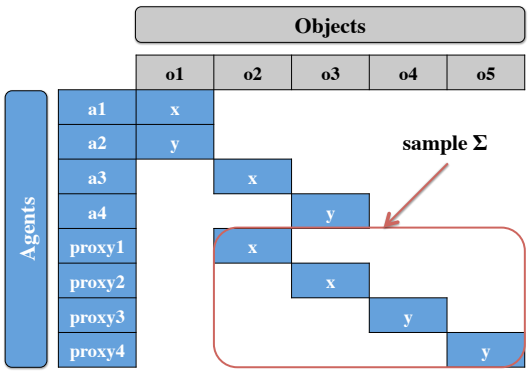

Figure 1: Minimal Peer Prediction with Private Priors.

Figure 1 shows one possible outcome of the elicitation process for a binary evaluation space $\{x, y\}$. To score agent a_1, the minimal peer prediction first builds Σ sample based on the reports of the proxy agents. Since Σ is double-mixed, the mechanism acquires the reports of agents a_3 and a_4, that, together with the report of agent a_2, define agent a_1's score. In this case, the (ex post) score of agent a_1 is equal to $\frac{1}{2}$.

Theorem 1. *The Minimal Peer Prediction with Private Priors is strictly Bayes-Nash incentive compatible whenever sample Σ contains at least two times more elements than the evaluation space $\{x, y, z, ...\}$, i.e. $|\Sigma| \geq 2|\{x, y, z, ...\}|$.*

Proof. Consider an agent a_1 whose evaluation is equal to x, and suppose other agents are honest, including an agent a_2 whose evaluation is equal to y. Due to the independence of $Y^{o_2}_{a_3}$ and $Y^{o_3}_{a_4}$ and linearity of expectations, the expected score of agent a_1 for reporting $\tilde{x}$ is equal to:

$$\bar{u}(\tilde{x}, y) = \pi(\Sigma)\left[\frac{1}{2} + Pr(y|\tilde{x}) - \frac{1}{2}\sum_z Pr(z|\tilde{x})^2\right]$$

where $\pi(\Sigma)$ is the probability that Σ is double-mixed. Fully mixed priors and $|\Sigma| \geq 2|\{x, y, z, ...\}|$ imply that $\pi(\Sigma) > 0$, so $\bar{u}(\tilde{x}, y)$ has the structure of the quadratic scoring rule, scaled by $\pi(\Sigma)$, that rewards agent a_1's posterior beliefs $Pr(\cdot|\tilde{x})$ with the realization of the outcome specified by agent a_2's report. Since the quadratic scoring rule is in expectation maximized when agent a_1 reports her true belief $Pr(\cdot|x)$, agent a_1 is incentivized to report honestly her evaluation, i.e. $\tilde{x} = x$. Moreover, agent a_1 is strictly incentivized to do so because of the stochastic relevance of her posterior beliefs. □

The mechanism requires $2|\{x, y, z, ...\}|$ statistically similar objects in addition to the object rated by agent a. This is usually not the issue since the answer space $\{x, y, z, ...\}$ is often significantly smaller than the amount of objects that the mechanism wants to evaluate. For example, Amazon's ratings consists of only 5 discrete elements (5 stars), while large number of products have statistically similar features.

Notice that the score $u(Y^{o_1}_{a_1}, Y^{o_1}_{a_2})$ takes values in $[0, \frac{3}{2}]$, so the payments are ex post individually rational and bounded. By multiplying it with a constant $\alpha > 0$ and adding a constant β, we can achieve that payoffs take values from an arbitrary interval. Furthermore, for N_a agents, $\alpha = \frac{2}{3}\frac{B}{N_a}$ and $\beta = 0$, the sum of all payoffs does not exceed a budget B.

The main drawback of the minimal peer prediction is that it assumes evaluations of different agents to be generated in a similar fashion. In the following section, we analyze a scenario that relaxes this assumption, and examine a protocol that elicits agents' beliefs along with their evaluations.

Heterogeneous Population

Unlike the previous section, a group of agents is now considered to have heterogeneous characteristics. More precisely, for two different agents a_1 and a_2, private signals for object o are obtained by sampling two different distributions, $Q(X^o_{a_1}|\Omega = \omega)$ and $Q(X^o_{a_2}|\Omega = \omega)$, respectively. This situation describes agents who form their opinions in different ways, e.g. an agent can be aware that her preferences significantly deviate from the rest of the population. In the general case, when two agents a_1 and a_2 can have arbitrary beliefs, it is not possible to create strict incentives that would elicit both agent a_1's and agent a_2's private signals.

Proposition 1. *No mechanism can provide incentives to heterogeneous agents that would make honest reporting a strict Bayes-Nash equilibrium.*

Proof (Sketch). Consider two agents a_1 and a_2 that evaluate object o. Assume agent a_1 believes that her peers a_3 sample from the distribution:

$$R_{a_1}(X^o_{a_3} = x|\Omega = \omega) = R_{a_1}(X^o_{a_1} = x|\Omega = \omega), \forall x$$

while agent a_2 believes that her peers a_4 sample from the distribution:

$$R_{a_2}(X^o_{a_4} = x|\Omega = \omega) = R_{a_2}(X^o_{a_2} = y|\Omega = \omega)$$
$$R_{a_2}(X^o_{a_4} = y|\Omega = \omega) = R_{a_2}(X^o_{a_2} = x|\Omega = \omega)$$
$$R_{a_2}(X^o_{a_4} = z|\Omega = \omega) = R_{a_2}(X^o_{a_2} = z|\Omega = \omega), \forall z \neq x, y$$

Notice that peers a_3 include agent a_2, peers a_4 include agent a_1, and peers a_3 and a_4 have common agents. What matters here, however, is that agents a_1 and a_2 have different beliefs regarding the private signals of their peers.

Furthermore, let $R_{a_1}(X^o_{a_1} = z|\Omega = \omega) = R_{a_2}(X^o_{a_2} = z|\Omega = \omega)$ for all evaluations z. Since $X^o_{a_3} = x$ from agent a_1's perspective has the same statistical properties as $X^o_{a_4} = y$ from agent a_2's perspective, the expected payoff of agent a_1 for reporting x is the same as the expected payoff of agent a_2 for reporting y when agents have equal evaluations $X^o_{a_1} = X^o_{a_2}$. Therefore, the mechanism cannot strictly incentivize the agents to report honestly both evaluations x and y. □

While truthfulness cannot be achieved in this scenario, one can incentivize an agent to report *consistently*, i.e. to report the same report for equal evaluations. More formally:

Definition 1. *An agent a's strategy of reporting her private signals (evaluations) is consistent if it can be described by a bijective function $\sigma : \{x, y, z, ...\} \to \{x, y, z, ...\}$ that maps evaluations to reports. That is, whenever agent a's evaluation is equal to x, her report is equal to $\sigma(x)$.*

Bayesian Truth Serum with Private Priors

We present a mechanism that is based on the principles of the Bayesian Truth Serum (BTS) (Prelec 2004), where an agent is rewarded for providing two reports: *information* report that corresponds to her private signal, and *prediction* report that corresponds to her posterior belief regarding the distribution of information reports. We develop on the divergence-based BTS from (Radanovic and Faltings 2014), which penalizes agents who have similar information reports, but significantly different prediction reports. In our setting we cannot utilize such comparison as the setting does not assume a common prior among agents. Instead, we incentivize an agent to be *consistent* by comparing her own reports for different objects.

In this section, we consider a slightly different scenario, in which a single agent a evaluates multiple objects, one after another, until a certain stopping criterion is reached. We model the criterion probabilistically by assuming that after each round the process terminates with a probability equal to $\tau \in (0, 1)$. That is, an agent at round i is asked to evaluate more objects with probability τ, or stated slightly differently, the number of objects to be evaluated by agent a is sampled from the geometric distribution with parameter τ, and this number is not announced to agent a.

Consider an agent a who evaluates objects $\{o_1, ..., o_m\}$ in a sequential manner, and is asked to provide for each of them: her evaluation $Y_a^{o_i}$ and her prediction $\mathbf{F}_a^{o_i}$ regarding how other agents have evaluated o_i. Our approach of eliciting consistent behaviour is based on linking an agent's evaluations Y to her predictions $\mathbf{F}$. Namely, an agent's predictions should be similar for similar evaluations, so we can use this fact to construct incentives that encourage consistent reporting. In order to apply such a principle, one needs to define a similarity measure of predictions.

Notice that each scoring rule is associated with a divergence function that measures the difference between the expected scores of the true (optimal) prediction $\mathbf{F}_{true}$ and the reported prediction $\mathbf{F}_{rep}$. For example, the divergence of the quadratic score (1) is equal to:

$$D(\mathbf{F}_{true}||\mathbf{F}_{rep}) = \frac{1}{2}\sum_z (\mathbf{F}_{true}(z) - \mathbf{F}_{rep}(z))^2 \qquad (2)$$

The divergence function (2) is, in fact, the Euclidean distance between two predictions, so it represents a suitable candidate for a similarity measure of predictions. As already mentioned, equal evaluations lead to similar predictions, which we can utilize to detect inconsistencies in reports, and, thus, penalize inconsistent behaviour. In particular, in our approach we penalize an agent if the divergence (2) between her two predictions that correspond to equal evaluations is larger than a random threshold.

Bayesian Truth Serum with Private Priors. The mechanism has the following structure:

1. Sample a number m from the geometric distribution with parameter $\tau \in (0, 1)$, and randomly select m a priori similar objects $\{o_1, ..., o_m\}$.[3]
2. Ask an agent a to rate objects $\{o_1, ..., o_m\}$ in a sequential manner - when rating an object o_i, the agent does not know if there are more objects to rate, she only knows that the probability of having additional objects is equal to τ.
3. For each object o_i agent a reports:
   - *information* report $Y_a^{o_i}$, i.e. her evaluation of the object;
   - *prediction* report $\mathbf{F}_a^{o_i}$, i.e. her prediction regarding the frequency of information reports for the object.
4. Agent a is rewarded with two scores: a prediction score, that she receives after rating each object, and an information score, that she receives after rating all the objects.
5. *Prediction* score u_1 rewards an agent a for her prediction report $\mathbf{F}_a^{o_i}$ using the quadratic scoring rule and the information report of a randomly chosen peer agent p who evaluated object o_i:
$$u_1 = S(\mathbf{F}_a^{o_i}, Y_p^{o_i})$$
6. *Information* score u_2 rewards an agent a for her information reports $\{Y_a^{o_1}, ..., Y_a^{o_m}\}$ using the fact that similar evaluations should lead to similar prediction reports. More precisely, the agent is penalized (not rewarded by 1) if the divergence (defined by (2)) between any two prediction reports whose corresponding information reports are equal is larger than a randomly chosen threshold $\theta \in (0, 1)$:
$$u_2 = \begin{cases} 0 & \text{if } \max_{Y_a^{o_i} = Y_a^{o_j}} D(F_a^{o_i}||F_a^{o_j}) > \theta \\ 1 & \text{otherwise} \end{cases}$$
If there are no two equal information reports $Y_a^{o_i} = Y_a^{o_j}$, we set the information score to $u_2 = 1$.

When rating object o_i, agent a does not know whether there are more objects to rate. This is important in order to ensure strict incentives for consistent reporting at each round of the mechanism. Namely, if agent a knows that object o_m, evaluated as y, is the last object to evaluate, and all objects prior to o_m are evaluated as x, then she is indifferent between reporting y and z. This is due to the fact that there does not exist any other object evaluated by agent a as y or z, so the information score is unaffected for reports y and z.

To illustrate the principle of the mechanism, consider an agent a who evaluates object o_1 as x. Her belief about the evaluations of other agents who evaluated the same object o_1 is $Pr(\cdot|x)$, where $\cdot$ denotes a possible evaluation. Thus, if she believes that others are honest, her best response is to report the prediction $\mathbf{F}_a^{o_1} = Pr(\cdot|x)$. Now, consider another object o_2, that agent a evaluates as y. Naturally, agent a's belief about the evaluations of other agents for object o_2 is equal to $Pr(\cdot|y)$, and her best response to truthful behaviour of others is to report the prediction $\mathbf{F}_a^{o_2} = Pr(\cdot|y)$. To maximize her information score, agent a should report different information reports for objects o_1 and o_2, because $D(\mathbf{F}_a^{o_1}||\mathbf{F}_a^{o_2}) > 0$ could be larger than random threshold $\theta > 0$. In other words, in order to avoid the information score equal to 0, agent a should report consistently.

Theorem 2. *Consider the Bayesian Truth Serum with Private Priors. The strictly best response of an agent a to truthful behaviour of the other agents is to report at each stage i*

[3] Number of tasks m can be controlled by both the total number of tasks M and parameter τ

her true prediction report $\mathbf{F}_a^{o_i} = Pr(X_p^{o_i}|X_a^{o_i})$ *and a consistent information report* $Y_a^{o_i} = \sigma(X_a^{o_i})$.

Proof. Consider an agent a and suppose all the other agents are honest. The maximal values of the prediction scores are obtained when the agent reports her true prediction reports. Due to the stochastic relevance, this is a strict optimum.

We also need to show that agent a is strictly incentivized to use a strategy σ when reporting her information reports, and that the information score does not change strict incentives for the prediction reports. It suffices to prove that when agent a reports her true predictions, the maximum of the information report is obtained when she uses σ reporting strategy. Suppose that agent a does report her honest predictions. If she uses σ strategy, her information score is indeed equal to 1, which is its optimal value. Namely, in that case we have $\max_{Y_a^{o_i}=Y_a^{o_j}} D(\mathbf{F}_a^{o_i}||\mathbf{F}_a^{o_j}) = 0 < \theta$. Therefore, the agent is incentivized to report honestly her prediction report and consistently her information report.

Lastly, we need to show that not using a bijective function σ only lowers the expected payoff of her information score. Namely, at each stage i, there is always a strictly positive probability p that agent a will experience all possible evaluations $\{x, y, z, ...\}$, and that threshold θ will be small enough so that $D(Pr(\cdot|X_a^{o_i} = x)||Pr(\cdot|X_a^{o_j} = y)) > \theta, \forall x \neq y$. Hence, her information score is in expectation less than or equal to: $(1-p)\cdot 1 + p\cdot 0 < 1$. Therefore, reporting honest predictions $F_a^{o_i}$ and consistent evaluations $\sigma(X_a^{o_i})$ is agent a's best response to honest behaviour of the other agents. □

The direct consequence of Theorem 2 is that truthful reporting is an equilibrium strategy.

Theorem 3. *Truthful reporting is a weak Perfect Bayes-Nash equilibrium in the Bayesian Truth Serum with Private Priors.*

Proof. The claim follows from Theorem 2 and the fact that σ can also be an identity function, i.e. $\sigma(x) = x, \forall x$. □

Note that among all bijective functions σ, the identity function corresponding to truthful reporting is the only one that can be implemented without coordination among agents and thus has the lowest cost, so that in practice the truthful equilibrium can be expected to be strict.

One might be worried that the number of tasks m evaluated by an agent a could potentially be large. However, the analysis of the mechanism is valid for any parameter $\tau \in (0, 1)$, so τ can be set to arbitrarily small value to make probability of having large number of tasks m very small.

Score u_1 of the the Bayesian Truth Serum with Private Priors is the quadratic scoring rule that takes values in $[0, 1]$, while score u_2 is either 0 or 1, so the mechanism is ex-post individually rational and provide bounded incentives. As shown in the previous section, this implies that the scores can be scaled so that participants receive positive payments, while the total payoff does not exceed a fixed budget.

As making and interpreting observations is costly, some agents can be expected to take shortcuts and submit reports that are independent of the object and chosen according to some distribution $rand$ (which may also be deterministic and always chose the same value). We call such a strategy heuristic reporting. Suppose that α fraction[4] of all agents do not respond to incentives, but instead they generate reports randomly according to a distribution $rand$. If the other $1-\alpha$ agents are honest, consistent reporting remains the best response. The intuition is that an agent's belief regarding the reports of other agents changes, but she is still incentivized to report consistently. Having in mind that truthfulness is a consistent behaviour, we obtain that truthful reporting remains to be an equilibrium strategy for agents who respond to the incentives.

Proposition 2. *Suppose that* $\alpha < 1$ *fraction of agents reports heuristically according to some random distribution* $rand$*. Then for the other agents, truthful reporting is a weak Perfect Bayes-Nash equilibrium of the Bayesian Truth Serum with Private Priors.*

Proof (Sketch). Consider an agent a and suppose there are α agents who report heuristically and $1 - \alpha$ agents who are honest. An agent a's posterior belief regarding her peer's report $Pr(Y_p^{o_i}|X_a^{o_i})$ is not anymore equal to $Pr(X_p^{o_i}|X_a^{o_i})$, so to avoid confusion let us denote $Pr(Y_p^{o_i} = y|X_a^{o_i} = x)$ by $\hat{Pr}(y|x)$, and $Pr(X_p^{o_i} = y|X_a^{o_i} = x)$, as usual, by $Pr(y|x)$. The connection between the two is $\tilde{Pr}(y|x) = (1 - \alpha) \cdot Pr(y|x) + \alpha \cdot rand(y)$. Since $\tilde{Pr}$ can be seen as a possible belief of agent a that has the same properties as the belief Pr, including stochastic relevance, the analysis from the proof of Theorem 2 applies here as well. Hence, the claim follows from Theorem 3. □

Conclusion

In this paper we designed two mechanisms that allow agents to have private prior beliefs. We showed that when agents' characteristics are homogeneous, there exists a simple and intuitive mechanism for truthful elicitation of agents' private information. On the other hand, for agents with heterogeneous characteristics, it is not possible to produce strict incentives for truthfulness. However, agents can be incentivized to report consistently, even in the presence of presence of agents who do not respond to the incentives. This further implies that truthful reporting is a (weak) equilibrium, but robust to the noise of heuristic reports. The most interesting direction for future work would be to analyze settings that lie in between the two explained in this paper, i.e. to explore settings that do not assume agents who have fully homogeneous characteristics, but do place some restrictions on how agents obtain their private information.

Acknowledgments

The work reported in this paper was supported by Nano-Tera.ch as part of the OpenSense2 project. We thank the anonymous reviewers for useful comments and feedback.

[4]More precisely, an agent does not respond to incentives with probability α.

References

Chen, Y., and Pennock, D. M. 2007. A utility framework for bounded-loss market makers. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI2007)*, 49–56.

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd ACM International World Wide Web Conference (WWW13)*.

Gneiting, T., and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102:359–378.

Goel, S.; Reeves, D. M.; and Pennock, D. M. 2009. Collective revelation: A mechanism for self-verified, weighted, and truthful predictions. In *Proceedings of the 10th ACM conference on Electronic commerce (EC 2009)*.

Hanson, R. D. 2003. Combinatorial information market design. *Information Systems Frontiers* 5(1):107–119.

Jurca, R., and Faltings, B. 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06)*, 190–199.

Jurca, R., and Faltings, B. 2007. Robust incentive-compatible feedback payments. In *Agent-Mediated Electronic Commerce*, volume LNAI 4452, 204–218. Springer-Verlag.

Jurca, R., and Faltings, B. 2011. Incentives for answering hypothetical questions. In *Workshop on Social Computing and User Generated Content (EC-11)*.

Lambert, N., and Shoham, Y. 2008. Truthful surveys. In *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE 2008)*.

Lambert, N., and Shoham, Y. 2009. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the tenth ACM conference on Electronic Commerce*, 109–118.

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

Prelec, D. 2004. A bayesian truth serum for subjective data. *Science* 34(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)*.

Radanovic, G., and Faltings, B. 2014. Incentives for truthful information elicitation of continuous signals. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14)*.

Riley, B. 2014. Minimum truth serums with optional predictions. In *Proceedings of the 4th Workshop on Social Computing and User Generated Content (SC14)*.

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66(336):783–801.

Waggoner, B., and Chen, Y. 2013. Information elicitation sans verification. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC13)*.

Waggoner, B., and Chen, Y. 2014. Output agreement mechanisms and common knowledge. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Witkowski, J., and Parkes, D. C. 2012a. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC' 12)*, 964–981.

Witkowski, J., and Parkes, D. C. 2012b. A robust bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*.

Witkowski, J., and Parkes, D. C. 2013. Learning the prior in minimal peer prediction. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC 2013)*.

Zhang, P., and Chen, Y. 2014. Elicitability and knowledge-free elicitation with peer prediction. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (AAMAS '14)*.