A Robust Bayesian Truth Serum for Small Populations (Technical Report)

Jens Witkowski
Department of Computer Science
Albert-Ludwigs-Universität
Freiburg, Germany
witkowsk@informatik.uni-freiburg.de

David C. Parkes
School of Engineering & Applied Sciences
Harvard University
Cambridge, MA, USA
parkes@eecs.harvard.edu

Abstract

Peer prediction methods allow the truthful elicitation of private signals (e.g., experiences, or opinions) in regard to a true world state when this ground truth is unobservable. The original peer prediction method is incentive compatible for any finite number of agents $n \geq 2$ but critically relies on a common prior, shared by all agents and the center. The Bayesian Truth Serum (BTS) relaxes this assumption. While it still assumes that the agents share a common prior, this prior need not be known by the center. However, BTS is proven to be incentive compatible only for a large enough number of agents, and this number depends on the prior and is thus unknown to the mechanism. In this paper, we present a robust BTS for the elicitation of binary information which is incentive compatible for any $n \geq 3$, taking advantage of a particularity of the quadratic scoring rule. Our mechanism is the first peer prediction method that does not rely on knowledge of the common prior to provide strict incentive compatibility for any $n \geq 3$. Moreover, and in contrast to the original BTS, our mechanism is numerically robust and ex post individually rational.

Introduction

Web services that are built around user-generated content are ubiquitous. Examples include reputation systems, where users leave feedback about the quality of products or services, and crowdsourcing platforms, where users (workers) are paid small rewards to do human computation tasks, such as annotating an image. Whereas statistical estimation techniques (Raykar et al. 2010) can be used to resolve noisy inputs, for example in order to determine the image tags most likely to be correct or the most likely true quality of a product or service, they are appropriate only when user inputs are informative in the first place. But what if providing accurate information is costly for users, or if users otherwise have an external incentive for submitting false inputs?

The peer prediction method (Miller, Resnick, and Zeckhauser 2005) addresses the quality control problem by providing payments (in cash, points or otherwise) that align an agent's own interest with providing inputs that are predictive of the inputs that will be provided by other agents. Formally, the peer prediction method provides strict incentives for providing truthful inputs (e.g., in regard to a user's information about the quality of a product, or view on the correct label for a training example) for a system of two or more agents, and when there is a common prior amongst agents and, critically, known to the mechanism.

The Bayesian Truth Serum (BTS) by Prelec (2004) still assumes that agents share a common prior, but does not require this to be known by the mechanism. In addition to an information report from an agent, BTS asks each agent for a prediction report, that reflects the agent's belief about the distribution of information reports in the population. An agent's payment depends on both reports, with an information component that rewards reports that are "surprisingly common," i.e. more common than collectively predicted, and a prediction component that rewards accurate predictions of the reports made by others. Compared to the original peer prediction method, a significant drawback of BTS in practice is that it only provides incentives for truthful reports for a large enough number of agents, and this number depends on the prior and is thus unknown to the mechanism. In addition, BTS may leave a participant with a negative payment and it is not numerically robust.

In this paper, we present a *robust Bayesian Truth Serum* (RBTS) for the elicitation of binary information which, to the best of our knowledge, is the first peer prediction method that does not rely on knowledge of the common prior to provide strict incentive compatibility for any number of agents $n \geq 3$. RBTS is also *ex post* individually rational (so that no agent makes a negative payment in any outcome) and numerically robust, being well defined for all possible agent reports. Moreover, the mechanism seems conceptually simpler than BTS and the incentive analysis is more straightforward. RBTS applies to the same setting and takes the same reports as BTS. As in BTS, an agent's payment consists of two components, one component that depends on an agent's information report and a second that depends on an agent's prediction report. The main innovation in RBTS is to induce a "shadow" posterior belief report for an agent i from her information report and the prediction report of another agent j, adjusting this prediction report in the direction suggested by agent i's information report. We couple this with a property of the quadratic scoring rule by which an agent prefers a shadow belief that is as close as possible to her true posterior. In order to determine the agent's payment, we then apply both the shadow belief and the agent's prediction report to the quadratic scoring rule with the information report of a third agent k as the event to be predicted.

In contrast to BTS, RBTS is defined here only for the case of binary information reports; e.g., good or bad experiences, or true or false classification labels. Many interesting applications involve binary information reports. This is supported by the fact that Prelec's own experimental papers have adopted the binary signal case (Prelec and Seung 2006; John, Loewenstein, and Prelec 2011). Also note that as the number of possible information reports increases, so does the difficulty imposed on users in providing the prediction report, which must expand to include estimates for the additional possible information reports. We leave to future research the study of extensions of RBTS that can incorporate more than two signals.

Related Work

In addition to the original peer prediction method and the original Bayesian Truth Serum, there is other related work:

Jurca and Faltings (2007) provide an extension to the original peer prediction method. While they assume a common prior shared by the agents and the center, the agents are allowed to have small deviations in regard to this prior. They show that there is a trade-off between the required budget and the robustness in regard to these deviations from the prior. The key difference to our work is that we do not assume any knowledge about the common prior on behalf of the center.

In another line of work, Jurca and Faltings (2008) propose *online* polling mechanisms, where the current empirical frequency of reports is published and updated as agents arrive. As in BTS and RBTS, the setting assumes a common prior known to the agents but unknown to the center. While their scheme only requires the information report (and not the prediction report), it is not incentive compatible in the sense of our work. Rather, agents must behave strategically in deciding how to report information to the mechanism (specifically, in selecting a "reference"). In this way, the scheme does not share the conceptual simplicity of our approach. One of their main criticisms of BTS is that it needs to withhold all information reports until the end of the poll. This criticism does not apply to RBTS, which easily adapts to online settings by sequentially scoring groups of three agents, and subsequently releasing their reports (which can be published as empirical frequencies). Jurca and Faltings (2011) also give an impossibility result in regard to incentive compatibility without a prior known to the center, but this precludes schemes such as BTS and RBTS, in which agents submit both an information report and a prediction report.

A setting similar to online polling is studied by Lambert and Shoham (2008), and in this case without even requiring a common prior to agents. However, their mechanism is only *weakly* incentive compatible, i.e. in the equilibrium, all agents are indifferent between being truthful and misreporting. The peer prediction mechanism with private beliefs (Witkowski and Parkes 2011) also studies a setting without a common prior to agents. They do achieve strict incentive compatibility for $n \geq 2$, but their mechanism critically relies on "temporal separation," i.e. the ability to elicit relevant information from an agent both before and after she receives her signal. While this is possible for settings such as product rating environments, it prevents the application to settings such as opinion polls, where an agent already holds her information when arriving to the mechanism.

The Setting

There are $n \geq 3$ rational, risk-neutral agents who seek to maximize their expected payment. They all share the same probabilistic belief system, which consists of two main elements: *types* and *signals*. The type T is a random variable which can adopt values in $\{1, \ldots, m\}$, $m \geq 2$ and represents the true state of the world. Each agent i observes a signal represented by random variable S_i that is binary and drawn from $\{0, 1\}$, sometimes represented $\{l, h\}$ and referred to as "low" and "high" respectively. The signal can be thought to represent an agent's experience or opinion. A generic signal is denoted by random variable S. The agents have common beliefs $\Pr(T = t)$ and $\Pr(S = h|T = t)$ on the conditional probability of observing a high signal given each possible state t. Collectively, we refer to the shared probabilistic belief system as the *common prior*.

Definition 1. *We require for a prior to be* admissible *that*

- *Every type occurs with positive probability, so that* $\Pr(T = t) > 0$ *for all* $t \in \{1, \ldots, m\}$.
- *Types are distinct, such that* $\Pr(S = h|T = t) \neq \Pr(S = h|T = t')$ *for any two* $t \neq t'$. *We adopt the convention that* $\Pr(S = h|T = 1) < \Pr(S = h|T = 2) < \ldots < \Pr(S = h|T = m)$.
- *The signal conditionals are fully mixed, with* $0 < \Pr(S = h|T = t) < 1$ *for all* t.

It bears emphasis that—with exception of its admissibility—neither BTS nor RBTS assume the center to have any knowledge about the prior.

Given an agent i's realized signal s_i, the agent can update her posterior belief $\Pr(S_j = h|S_i = s_i)$ about the probability of another agent j receiving a high signal. Because of the common prior, we can denote a generic agent's posterior following a high and a low signal with $p_{\{h\}} = \Pr(S_j = h|S_i = h)$ and $p_{\{l\}} = Pr(S_j = h|S_i = l)$, respectively. We refer to these as "first order" signal posteriors and we have:

$$\Pr(S_j = h|S_i = s_i) = \sum_{t=1}^{m} \Pr(S_j = h|\, T = t) \cdot \Pr(T = t|\, S_i = s_i), \quad (1)$$

where the posterior on type can be determined in the usual way from Bayes' rule, being equal to

$$\Pr(T = t|\, S_i = s_i) = \frac{\Pr(S_i = s_i|\, T = t)\Pr(T = t)}{\Pr(S_i = s_i)}, \quad (2)$$

and the denominator being

$$\Pr(S_i = s_i) = \sum_{t=1}^{m} \Pr(S_i = s_i|\, T = t) \cdot \Pr(T = t). \quad (3)$$

These signal posteriors can be computed analogously in the case where an agent has knowledge of two signals. We extend the notation, so that $p_{\{h,l\}}$ represents this "second-order" posterior following knowledge of a high signal and a low signal. For agent i in particular, we have $p_{\{h,l\}} = \Pr(S_k = h | S_i = h, S_j = l)$ for any distinct $j, k \neq i$. In this case, agent i first updates the posterior on type T, $Pr(T = t|S_i = s_i)$, which becomes the *a priori* belief on type for the purpose of doing a second round of Bayesian updates.

The Bayesian Truth Serum

In this section, we explain the original Bayesian Truth Serum (BTS) by Prelec (2004).[1] While we present the binary version of this method, BTS is defined for an arbitrary number of signals. Note that we adopt the convention $\{0, 1\}$ for signals to help with the presentation.

First, every agent i is asked for two reports:

- **Information report**: Let $x_i \in \{0, 1\}$ be agent i's reported signal.
- **Prediction report**: Let $y_i \in [0, 1]$ be agent i's report about the frequency of high signals in the population.

The scoring of agent i then involves three steps:

1. For every agent $j \neq i$, calculate the arithmetic mean of all agent's signal reports except i and j (with Laplacian smoothing to avoid infinite scores associated with zero frequencies):

$$\bar{x}_{-ij} = \frac{1}{n}\left(\sum_{k\neq i,j} x_k + 1\right) \tag{4}$$

2. For every agent $j \neq i$, calculate the geometric mean of all agent's predictions in regard to both high and low signals, except i and j:

$$\bar{y}_{-ij} = \left(\prod_{k\neq i,j} y_k\right)^{\frac{1}{n-2}}, \quad \bar{y}'_{-ij} = \left(\prod_{k\neq i,j} (1-y_k)\right)^{\frac{1}{n-2}} \tag{5}$$

3. Calculate the BTS score for agent i:

$$\begin{aligned} u_i = & \underbrace{\sum_{j\neq i}\left(x_i \ln\frac{\bar{x}_{-ij}}{\bar{y}_{-ij}} + (1-x_i)\ln\frac{1-\bar{x}_{-ij}}{\bar{y}'_{-ij}}\right)}_{\text{information score}} \\ & + \underbrace{\sum_{j\neq i}\left(\bar{x}_{-ij}\ln\frac{y_i}{\bar{x}_{-ij}} + (1-\bar{x}_{-ij})\ln\frac{1-y_i}{1-\bar{x}_{-ij}}\right)}_{\text{prediction score}} \end{aligned} \tag{6}$$

For the case of $n \to \infty$, this simplifies and the summation over $j \neq i$ in Equation 6 can be replaced with the information score and prediction score computed for just one, randomly selected, $j \neq i$.

Properties

A Bayesian Truth Serum mechanism is *Bayes-Nash incentive compatible* if it is a strict Bayes-Nash equilibrium for all agents to (1) report their true signal and (2) predict that the frequency of high signals in the population is that of their signal posterior.

Theorem 1. *(Prelec 2004) The Bayesian Truth Serum is Bayes-Nash incentive compatible for $n \to \infty$ and all admissible priors.*

Prelec comments that the result also holds for suitably large, finite n with the actual threshold depending on the common prior. However, BTS does not align incentives for small groups of agents. Moreover, it does not satisfy participation constraints in that it is not *interim* individually rational (interim IR) for small groups, meaning that an agent's expected payment can be negative.

Theorem 2. *The Bayesian Truth Serum is not Bayes-Nash incentive compatible or interim IR for $n = 3$.*

Certainly this can be understood from Prelec's treatment of BTS. Note, however, that BTS has been discussed in various places (e.g., Jurca and Faltings, 2008; Chen and Pennock, 2010) without noting this important caveat. For this reason, we provide a constructive example, which serves to highlight the difference between the $n \to \infty$ and the small n case. The example is not unique and it does not rely on $n = 3$. Generally the number of agents required for BTS to be Bayes-Nash incentive compatible depends on the prior and is hard to characterize.

Example 1 (BTS and $n = 3$). Consider three agents sharing the following prior with $m = 2$: $\Pr(T = 2) = 0.7, \Pr(S = h|T = 2) = 0.8$ and $\Pr(S = h|T = 1) = 0.1$. Based on this, the posterior signal beliefs (following Bayes' rule) are $p_{\{h\}} = \Pr(S_j = h|S_i = h) = 0.764$ and $p_{\{l\}} = \Pr(S_j = h|S_i = l) = 0.339$.

Consider agent 1, and assume agents 2 and 3 are truthful. Assume that $S_1 = h$, so that agent 1's truthful reports are $x_1 = 1$ and $y_1 = 0.764$. The expected score for the terms in Equation 6 that correspond to agent $j = 2$ when agent 1 reports truthfully is:

$$E\left[\ln\frac{\bar{X}_{-13}}{\bar{Y}_{-13}} + \bar{X}_{-13}\ln\frac{0.764}{\bar{X}_{-13}} + (1–\bar{X}_{-13})\ln\frac{1-0.764}{1–\bar{X}_{-13}}\right],$$

with the expectation taken with respect to random variables $\bar{X}_{-13}$ and $\bar{Y}_{-13}$. With probability $p_{\{h\}} = 0.764$, agent 1 believes that $\bar{x}_{-13} = (1+1)/3 = 2/3$ and $\bar{y}_{-13} = 0.764$ and with probability $1 - p_{\{h\}} = 0.236$ that $\bar{x}_{-13} = (0+1)/3 = 1/3$ and $\bar{y}_{-13} = 0.236$.

We have expected *information score* $0.764 \ln \frac{2/3}{0.764} + 0.236 \ln \frac{1/3}{0.339} = -0.108$ and expected *prediction score* $0.764\left((2/3)\ln\frac{0.764}{2/3} + (1/3)\ln\frac{0.236}{1/3}\right) + 0.236\left((1/3)\ln\frac{0.764}{1/3} + (2/3)\ln\frac{0.236}{2/3}\right) = -0.117$, giving an expected score of -0.225. Considering also the score due to the $j = 3$ terms, the total expected score when agent 1 is truthful is -0.450.

[1] In his original paper, Prelec presents two versions of BTS, one for an infinite number of agents $n \to \infty$ and one for finite n. Given the focus of our paper, we present the latter version.

On the other hand, if agent 1 misreports and $x_1 = 0$, while still reporting $y_1 = 0.764$, then the expected *information score* component (for the $j = 2$ terms) would become, $E\left[\ln \frac{1-\bar{X}_{-13}}{\bar{Y}'_{-13}}\right] = 0.764 \ln \frac{1/3}{0.236} + 0.236 \ln \frac{2/3}{0.661} = 0.266$, which combines with the prediction score to give 0.149, and thus, considering also the $j = 3$ terms in Equation 6, yields a total expected score of 0.298. We see that agent 1 can do better by making a misreport and that BTS fails interim IR.

Example 2 (BTS and $n \to \infty$). Consider the same prior but now a large number of agents. In the limit, and with respect to the beliefs of agent 1, random variable $\overline{X}_{-ij}$ takes on value $\lim_{n\to\infty} \frac{1}{n}\left((n-2)\cdot p_{\{h\}} + 1\right) = p_{\{h\}}$ with probability 1; similarly, random variable $\overline{Y}_{-ij}$ takes on value $\lim_{n\to\infty}\left(p_{\{h\}}^{(n-2)p_{\{h\}}} \cdot p_{\{l\}}^{(n-2)(1-p_{\{h\}})}\right)^{1/(n-2)} = p_{\{h\}}^{p_{\{h\}}} \cdot p_{\{l\}}^{1-p_{\{h\}}} = 0.631$ with probability 1. Similarly, $\bar{Y}'_{-13}$ takes on value $(1-p_{\{h\}})^{p_{\{h\}}} \cdot (1-p_{\{l\}})^{1-p_{\{h\}}} = 0.301$ with probability 1. Putting this together, if agent 1 truthfully reports $x_1 = 1$ and $y_1 = 0.764$, her expected information score is $\ln \frac{0.764}{0.631} = 0.191$, and her expected prediction score is $0.764 \ln \frac{0.764}{0.764} + (1 - 0.764) \ln \frac{1-0.764}{1-0.764} = 0$, i.e. 0.191 in total. A misreport of $x_1 = 0$ gives expected information score of $\ln \frac{0.236}{0.301} = -0.243$. This confirms that BTS is Bayes-Nash incentive compatible in the large n limit.

Having demonstrated the failure of incentive alignment and interim IR in BTS, we also make the following observation in regard to its numerical robustness:

Proposition 3. *The score in the Bayesian Truth Serum is unboundedly negative for posterior reports $y_i \in \{0, 1\}$.*

Robust Bayesian Truth Serum

In this section, we introduce our own mechanism, the Robust Bayesian Truth Serum (RBTS). RBTS is incentive compatible for any $n \geq 3$, *ex post* individually rational, and numerically robust. Both the setting and the reports are identical to that of the original Bayesian Truth Serum (BTS).

Proper Scoring Rules

Proper scoring rules are functions that can be used to incentivize rational agents to truthfully announce their private beliefs about the likelihood of a future event.

Definition 2 (Binary Scoring Rule). *Given possible outcomes $\Omega = \{0, 1\}$, and a report $y \in [0, 1]$ in regard to the probability of outcome $\omega = 1$, a binary scoring rule $R(y, \omega) \in \mathbb{R}$ assigns a score based on report y and the outcome ω that occurs.*

First, the agent is asked for her belief report $y \in [0, 1]$. Second, an event $\omega \in \{0, 1\}$ materializes (observed by the mechanism) and, third, the agent receives payment $R(y, \omega)$.

Definition 3 (Strictly Proper Scoring Rule). *A binary scoring rule is* proper *if it leads to an agent maximizing her expected score by truthfully reporting her belief $p \in [0, 1]$ and* strictly proper *if the truthful report is the only report that maximizes the agent's expected score.*

An example of a strictly proper scoring rule is the binary quadratic scoring rule R_q, normalized her to give scores between 0 and 1:

$$\begin{aligned} R_q(y, \omega = 1) &= 2y - y^2 \\ R_q(y, \omega = 0) &= 1 - y^2. \end{aligned} \tag{7}$$

Proposition 4. *(Selten 1998) The binary quadratic scoring rule R_q is strictly proper.*

Note that if one applies a positive-affine transformation to a proper scoring rule, the rule is still proper. For a more detailed discussion of proper scoring rules in general, we refer to the article by Gneiting and Raftery (2007).

The RBTS Mechanism

First, every agent i is asked for two reports:

- **Information report**: Let $x_i \in \{0, 1\}$ be agent i's reported signal.
- **Prediction report**: Let $y_i \in [0, 1]$ be agent i's report about the frequency of high signals in the population.

In a second step, for each agent i, select two other agents $j = i + 1$ (modulo n) and $k = i + 2$ (modulo n), and calculate

$$y'_i = \begin{cases} y_j + \delta, & \text{if} \quad x_i = 1 \\ y_j - \delta, & \text{if} \quad x_i = 0 \end{cases}$$

where $\delta = \frac{1}{2}\min(y_j, 1 - y_j)$. The RBTS score for agent i is:

$$u_i = \underbrace{R_q(y'_i, x_k)}_{\text{information score}} + \underbrace{R_q(y_i, x_k)}_{\text{prediction score}} \tag{8}$$

Example 3 (RBTS and $n = 3$.) We illustrate RBTS with the numbers from Example 1, so that $p_{\{h\}} = 0.764$ and $p_{\{l\}} = 0.339$. In addition, we note that $p_{\{h,h\}} = 0.795$ and $p_{\{h,l\}} = 0.664$. We consider the perspective of agent 1 (as agent i) and let 2 and 3 play the roles of j and k, respectively. Throughout we assume agents 2 and 3 are truthful.

We first exemplify the computation of the mechanism in the concrete instance where $S_1 = h$, $S_2 = l$, and $S_3 = h$. Consider agent 1. For the information score, since $y_2 = 0.339$, we have $\delta = 0.1695$ and $y'_1 = y_2 + \delta = 0.339 + 0.1695 = 0.5085$. Since $x_3 = 1$, agent 1's information score is $2y'_1 - {y'_1}^2 = 2 \cdot 0.5085 - 0.5085^2 = 0.758$. Since $y_1 = 0.764$ and $x_3 = 1$, the prediction score is $2 \cdot 0.764 - 0.764^2 = 0.944$. In total, the agent's score is 1.703.

To establish that, when $S_1 = h$, agent 1 is best off reporting truthfully for the example prior, we need to consider the expected score and thus the distribution on possible signals of agents 2 and 3. For the prediction report, we have truthfulness because scoring rule $R_q(y_1, x_3)$ is strictly proper. In particular, agent 1's expected *prediction score* is $0.764 \cdot (2 \cdot 0.764 - 0.764^2) + 0.236 \cdot (2 \cdot 0.236 - 0.236^2) = 0.820$. For the expected *information score*, first consider truthful report $x_1 = 1$. In this case, y'_1 is adjusted upwards from the realized prediction report of agent 2. The expected information score of agent 1 is:

$$
\begin{aligned}
&\Pr(S_2 = h|S_1 = h)\cdot \\
&\quad\Big[\ \Pr(S_3 = h|S_1 = h, S_2 = h)\cdot R_q(0.764 + 0.118, 1) \\
&\quad\ + \Pr(S_3 = l|S_1 = h, S_2 = h)\cdot R_q(0.764 + 0.118, 0)\Big] \\
&+ \Pr(S_2 = l|S_1 = h)\cdot \\
&\quad\Big[\ \Pr(S_3 = h|S_1 = h, S_2 = l)\cdot R_q(0.339 + 0.1695, 1) \\
&\quad\ + \Pr(S_3 = l|S_1 = h, S_2 = l)\cdot R_q(0.339 + 0.1695, 0)\Big] \\
=\ & p_{\{h\}}\cdot\big[\ p_{\{h,h\}}\cdot\left(2\cdot 0.882 - 0.882^2\right) \\
&\quad + \left(1 - p_{\{h,h\}}\right)\cdot\left(1 - 0.882^2\right)\big] \\
&+ \left(1 - p_{\{h\}}\right)\cdot\big[p_{\{h,l\}}\cdot\left(2\cdot 0.5085 - 0.5085^2\right) \\
&\quad + \left(1 - p_{\{h,l\}}\right)\cdot\left(1 - 0.5085^2\right)\big] = 0.811
\end{aligned}
$$

For a report of $x_1 = 0$, agent 1's expected information score is:

$$
\begin{aligned}
&p_{\{h\}}\cdot\big[\ p_{\{h,h\}}\cdot R_q(0.764 - 0.118, 1) \\
&\quad + \left(1 - p_{\{h,h\}}\right)\cdot R_q(0.764 - 0.118, 0)\big] \\
&+ \left(1 - p_{\{h\}}\right)\big[\ p_{\{h,l\}}\cdot R_q(0.339 - 0.1695, 1) \\
&\quad + \left(1 - p_{\{h,l\}}\right)\cdot R_q(0.339 - 0.1695, 0)\big] = 0.748
\end{aligned}
$$

Agent 1 thus maximizes the expected information score by reporting her signal truthfully.

Note that RBTS is strictly Bayes-Nash incentive compatible for any $n \geq 3$ and any admissible prior. We go on to prove this in the following section.

Incentive Compatibility

We first establish some lemmas that are important in proving the Bayes-Nash incentive compatibility of RBTS. Note that Lemma 5 also establishes *stochastic relevance*, so that the signal posteriors are distinct for distinct signal observations.

Lemma 5. *It holds that* $1 > p_{\{h\}} > \Pr(S_j = h) > p_{\{l\}} > 0$ *for all admissible priors.*

Proof. To show is that $1 > \Pr(S_j = h|S_i = h) > \Pr(S_j = h) > \Pr(S_j = h|S_i = l) > 0$. The fully mixed property of admissible priors suffices to ensure that beliefs are always interior. Given the sorting property of admissible priors and Equation 1 it is then sufficient to show, for all $t' < m$, the following dominance condition on beliefs in regard to type:

$$
\begin{aligned}
\sum_{t=1}^{t'}\Pr(T = t|S_i = h) &< \sum_{t=1}^{t'}\Pr(T = t) \\
&< \sum_{t=1}^{t'}\Pr(T = t|S_i = l) \qquad (9)
\end{aligned}
$$

To see this, consider for example why dominance is sufficient for $p_{\{h\}} > p_{\{l\}}$. For contradiction, suppose Equation 9 without this strict inequality on signal posterior. Consider now a sequence of adjustments to the type posterior $\Pr(T|S_i = h)$ where in each step, working from $t' = m$ to 2, the probability assigned to $\Pr(T = t'|S_i = h)$ is reduced by the amount by which $\sum_{t=1}^{t'-1}\Pr(T = t|S_i = h)$ is less than $\sum_{t=1}^{t'-1}\Pr(T = t|S_i = l)$, with this probability assigned to the next lower type $\Pr(T = t' - 1|S_i = h)$. This operation maintains the invariant that the total probability from 1 to t' given h in the perturbed type posterior and in $\Pr(T = t|S_i = l)$ is equal, and thus the step is always feasible. Moreover, it moves probability from type t' to type $t' - 1$ and this decreases the signal posterior by the sorting property and Equation 1. Eventually, we have walked to posterior distribution $\Pr(T = t|S_i = l)$ (and thus signal posterior $p_{\{l\}}$ through a sequence of steps each of which strictly decreases the signal posterior. Contradiction, and given Equation 9 we must have $p_{\{h\}} > p_{\{l\}}$.

For the dominance condition, note that

$$
\begin{aligned}
\Pr(T = t|S_i = h) &\propto \Pr(S_i = h|T = t)\Pr(T = t) \quad (10) \\
\Pr(T = t|S_i = l) &\propto \Pr(S_i = l|T = t)\Pr(T = t) \quad (11)
\end{aligned}
$$

From this, and given that the type posterior probabilities will be normalized to sum to 1, it is sufficient for Equation 9 that we have

$$
\frac{\sum_{t=1}^{t'}\Pr(S_i = h|T = t)}{\sum_{t=1}^{m}\Pr(S_i = h|T = t)} < \frac{t'}{m} < \frac{\sum_{t=1}^{t'}\Pr(S_i = l|T = t)}{\sum_{t=1}^{m}\Pr(S_i = l|T = t)} \qquad (12)
$$

for all $t' < m$.

Recall that by the sorting property we have $\Pr(S_i = h|T = 1) < \Pr(S_i = h|T = 2) < \ldots < \Pr(S_i = h|T = m)$ and also $\Pr(S_i = l|T = 1) > \Pr(S_i = l|T = 2) > \ldots > \Pr(S_i = l|T = m)$. Now, for $t' = 1$, we have

$$
\begin{aligned}
\frac{\Pr(S_i = h|T = 1)}{\sum_{t=1}^{m}\Pr(S_i = h|T = t)} &= \frac{A}{mA + \epsilon_A} < \frac{1}{m} \\
< \frac{B}{mB - \epsilon_B} &= \frac{\Pr(S_i = l|T = 1)}{\sum_{t=1}^{m}\Pr(S_i = l|T = t)}, \qquad (13)
\end{aligned}
$$

where $A = \Pr(S_i = h|T = 1)$, $B = \Pr(S = l|T = 1)$ and $\epsilon_A, \epsilon_B > 0$. The first equality follows since the terms in the denominator are strictly increasing (by sorting), the second and third inequalities by algebra, and the second equality since the terms in the denominator are strictly decreasing.

For the case of $1 < t' < m$, we have

$$
\begin{aligned}
&\frac{\sum\limits_{t=1}^{t'}\Pr(S_i = h|T = t)}{\sum\limits_{t=1}^{m}\Pr(S_i = h|T = t)} < \frac{t'A}{t'A + \sum\limits_{t=t'+1}^{m}\Pr(S_i = h|T = t)} \\
&= \frac{t'A}{mA + \epsilon_A} < \frac{t'}{m} < \frac{t'B}{mB - \epsilon_B} \\
&= \frac{t'B}{t'B + \sum\limits_{t=t'+1}^{m}\Pr(S_i = l|T = t)} < \frac{\sum\limits_{t=1}^{t'}\Pr(S_i = l|T = t)}{\sum\limits_{t=1}^{m}\Pr(S_i = l|T = t)}, \qquad (14)
\end{aligned}
$$

where the first inequality follows by algebra, with $A = \Pr(S_i = h|T = t')$ and replacing smaller terms in both the numerator and denominator with A, the first equality recognizes that the remaining terms in the denominator are strictly

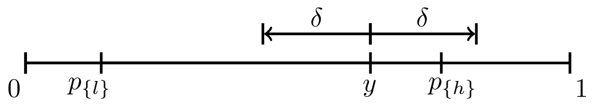

Figure 1: An example for the shadowing method with $y \in (p_{\{l\}}, p_{\{h\}})$. Note that $p_{\{l\}}$ is closer to $y_i' = y - \delta$ than to $y_i' = y + \delta$, and that $p_{\{h\}}$ is closer to $y_i' = y + \delta$ than to $y_i' = y - \delta$.

increasing, the second equality recognizes that the remaining terms in the denominator are strictly decreasing, and the final inequality follows by algebra, with $B = \Pr(S_i = l|T = t')$ and replacing larger terms in the numerator and denominator with B. This completes the proof. □

The following lemma extends Lemma 5 to second-order posteriors.

Lemma 6. *It holds that* $1 > p_{\{h,h\}} > p_{\{h\}} > p_{\{h,l\}} = p_{\{l,h\}} > p_{\{l\}} > p_{\{l,l\}} > 0$ *for all admissible priors.*

Proof. (Sketch) Consider $p_{\{h,h\}} > p_{\{h\}} > p_{\{h,l\}}$. This follows immediately from the same analysis as Lemma 5, with the type posterior $\Pr(T = t|S_i = h)$ taking the role of the a priori type belief $\Pr(T = t)$ in the analysis. Then, we have $p_{\{h,l\}} = p_{\{l,h\}}$, and the other case, $p_{\{l,h\}} > p_{\{l\}} = p_{\{l,l\}}$ can be shown analogously. □

Lemma 7 (Minimize Distance)**.** *Let* $p \in [0, 1]$ *be an agent's true belief about a binary future event. If the center scores the agent's belief report according to the quadratic scoring rule* R_q *but restricts the set of allowed reports to* $Y \subseteq [0, 1]$*, a rational agent will report the* $y \in Y$ *with minimal absolute difference* $|y - p|$*.*

Proof. The expected score of reporting y if p is the true belief is $E[y] = p \cdot \left(2y - y^2\right) + (1-p) \cdot \left(1 - y^2\right)$. The expected loss is thus $E[p] - E[y] = p \cdot \left(2p - p^2\right) + (1-p) \cdot \left(1 - p^2\right) - p \cdot \left(2y - y^2\right) - (1-p) \cdot \left(1 - y^2\right) = (p - y)^2$. That is, given a set of reports Y, a rational, selfish agent will report the y that minimizes $(p - y)^2$ and thus minimizes $|p - y|$. □

This property is not satisfied by all proper scoring rules. In particular, the two other frequently cited proper scoring rules, the logarithmic and the spherical rule, do not satisfy this property.

A Proper Scoring Rule for Eliciting Signals: The "Shadowing" Method

Proper scoring rules allow us to elicit probabilistic *beliefs*, but it is unclear how to elicit *signals* truthfully. The following "shadowing" method achieves just that.

Let $\omega \in \{0, 1\}$ denote a binary future event. (In the context of RBTS this will be the information report by some agent $k \neq i$.) In describing the method, we make this general by allowing agent i to have observed a sequence of signals $\mathcal{I} \in \{0, 1\}^o$ before with o denoting the number of signals agent i has observed.

1. Agent i receives a signal $S_i \in \{0, 1\} = \{l, h\}$ and, depending on this signal and previously observed signals $\mathcal{I}$, forms a posterior belief $p \in \{p_{\{l,\mathcal{I}\}}, p_{\{h,\mathcal{I}\}}\}$ about ω. (If the prior is admissible, it holds that $p_{\{l,\mathcal{I}\}} = Pr(\omega = 1|S_i = l, \mathcal{I}) < Pr(\omega = 1|S_i = h, \mathcal{I}) = p_{\{h,\mathcal{I}\}}$.)
2. The center asks the agent for signal report $x_i \in \{0, 1\}$ and transforms it into a probabilistic "shadow" posterior y_i' by:
$$y_i' = \begin{cases} y + \delta, & \text{if} \quad x_i = 1 \\ y - \delta, & \text{if} \quad x_i = 0, \end{cases} \tag{15}$$
where $y \in [0, 1]$ is a parameter of the method, and $\delta = \frac{1}{2}\min(y, 1 - y)$.
3. The "shadow" posterior report y_i' and the event ω that eventually materializes is then applied to the quadratic scoring rule R_q to give agent i a score of:
$$R_q(y_i', \omega) \tag{16}$$

Lemma 8 (Strict Properness)**.** *Agent* i *uniquely maximizes her expected score in the shadowing method by truthfully reporting her signal if* $y \in (p_{\{l,\mathcal{I}\}}, p_{\{h,\mathcal{I}\}})$*.*

Proof. Note that $0 < y < 1$ and thus $\delta > 0$. Without loss of generality, suppose agent i's signal is $S_i = h$ and signal posterior is $p_{\{h,\mathcal{I}\}}$. The argument is symmetric for $S_i = l$ and posterior $p_{\{l,\mathcal{I}\}}$. There are two cases:

- $y + \delta \leq p_{\{h,\mathcal{I}\}}$. But now $\delta > 0$, and so $y - \delta < y + \delta \leq p_{\{h,\mathcal{I}\}}$ and the result follows by Lemma 7.
- $y + \delta > p_{\{h,\mathcal{I}\}}$. But now $y < p_{\{h,\mathcal{I}\}}$ and so $(y + \delta) - p_{\{h,\mathcal{I}\}} < p_{\{h,\mathcal{I}\}} - (y - \delta)$ and the result follows by Lemma 7.

□

Theorem 9. *The Robust Bayesian Truth Serum is Bayes-Nash incentive compatible for any* $n \geq 3$ *and all admissible priors.*

Proof. Fix some i, j and k. It needs to be shown that given agents j and k report honestly, it is the unique best response for agent i to report honestly as well. Noting that the only effect of y_i is on the prediction score, and that strict Bayes-Nash incentive compatibility follows there from the use of a quadratic scoring rule, we then focus on x_i, which affects y_i' and thus the information score.

There are two cases to consider in regard to agent j:

1. $S_j = h$ and so $y_j = p_{\{h\}}$ in equilibrium. Conditioned on this additional signal information, agent i's posterior signal belief would be $p_{\{h,h\}}$ if $S_i = h$ and $p_{\{l,h\}}$ if $S_i = l$. By Lemma 8 it is sufficient that $p_{\{l,h\}} < y_j = p_{\{h\}} < p_{\{h,h\}}$, which holds by Lemma 6 and the fact that the prior is admissible.
2. $S_j = l$ and so $y_j = p_{\{l\}}$ in equilibrium. Conditioned on this additional signal information, agent i's posterior signal belief would be $p_{\{h,l\}}$ if $S_i = h$ and $p_{\{l,l\}}$ if $S_i = l$. By Lemma 8 it is sufficient that $p_{\{l,l\}} < y_j = p_{\{l\}} < p_{\{h,l\}}$, which holds by Lemma 6 and the fact that the prior is admissible.

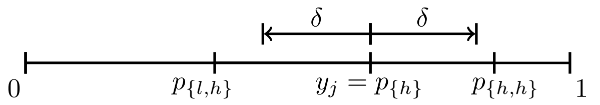

Figure 2: An example for RBTS in the $S_j = h$ case. Note that y_j is always strictly in between agent i's two possible second-order posteriors $p_{\{l,h\}}$ and $p_{\{h,h\}}$.

□

Other Properties & Discussion

Theorem 10. *The Robust Bayesian Truth Serum is ex post individually rational.*

Proof. The quadratic scoring rule is normalized to have scores on $[0, 1]$. □

Proposition 11. *The scores in the Robust Bayesian Truth Serum are in* $[0, 2]$ *for any reports from agents including any* $y_i \in [0, 1]$.

Proof. The binary quadratic scoring rule $R_q(y, \omega)$ is well-defined for any input $y \in [0, 1]$ and $\omega \in \{0, 1\}$. The inputs to R_q for computing the information score are $y := y'_i \in [0, 1]$ and $\omega := x_k \in \{0, 1\}$. Note that reports $y_j = 0$ and $y_j = 1$, in particular, lead to $y'_i = 0$ and $y'_i = 1$, respectively, which are well-defined inputs to R_q. The inputs for computing the prediction score are $y := y_i \in [0, 1]$ and $\omega := x_k \in \{0, 1\}$. □

Note also that if a designer has a particular budget $B > 0$ then a straightforward extension is to multiply R_q with a positive scalar $\alpha > 0$ to implement a mechanism that conforms with any budget constraint, since the total ex post cost is upper-bounded by $2\alpha n$.

A simple randomized extension of RBTS achieves constant *ex post* budget of $B > 0$ for groups of $n \geq 4$ by randomly excluding an agent from the population, running RBTS with budget $B > 0$ on the remaining $n - 1$ agents, and redistributing whatever remains from B to the excluded agent. Note that this extension to RBTS is still incentive compatible when the agents do not know which of them is the excluded agent. Moreover, the same technique can be used to implement a mechanism with $B = 0$.

Also note that in contrast to BTS, RBTS easily adapts to *online* polling settings, where the center seeks to publish partial information as agents arrive. Since RBTS achieves incentive compatibility for any group with $n \geq 3$ agents, the center can sequentially score groups of three, and subsequently release their reports.

Conclusion

In this paper, we introduced a novel Bayesian Truth Serum which takes the same inputs as the original Bayesian Truth Serum by Prelec but achieves strict Bayes-Nash incentive compatibility for any number of agents $n \geq 3$. It is interesting to see that a particularity of the quadratic scoring rule allows the development of proper scoring rule based mechanisms for eliciting *signals*. Using this "shadowing" method, we developed a constructive proof for the incentive compatibility of our Robust Bayesian Truth Serum. The quadratic scoring rule also proved to be advantageous in regard to numerical stability, because in contrast to the logarithmic rule, for example, it allows, and can adequately score, belief reports of 0. We believe that RBTS can have practical impact, providing a more principled approach to incentivize small groups of workers on crowdsourcing platforms such as Amazon Mechanical Turk (AMT), where the original Bayesian Truth Serum has already been shown to be useful for quality control (Shaw, Horton, and Chen 2011).

References

Chen, Y., and Pennock, D. M. 2010. Designing Markets for Prediction. *AI Magazine* 31:42–52.

Gneiting, T., and Raftery, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102:359–378.

John, L. K.; Loewenstein, G.; and Prelec, D. 2011. Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling. *Psychological Science*. to appear.

Jurca, R., and Faltings, B. 2007. Robust Incentive-Compatible Feedback Payments. In *Trust, Reputation and Security: Theories and Practice*, volume 4452 of *LNAI*. Springer-Verlag. 204–218.

Jurca, R., and Faltings, B. 2008. Incentives for Expressing Opinions in Online Polls. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08)*.

Jurca, R., and Faltings, B. 2011. Incentives for Answering Hypothetical Questions. In *Proceedings of the 1st Workshop on Social Computing and User Generated Content (SC'11)*.

Lambert, N., and Shoham, Y. 2008. Truthful Surveys. In *Proceedings of the 4th International Workshop on Internet and Network Economics (WINE '08)*, 154–165.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51(9):1359–1373.

Prelec, D., and Seung, S. 2006. An algorithm that finds truth even if most people are wrong. Working Paper.

Prelec, D. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306(5695):462–466.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning From Crowds. *Journal of Machine Learning Research* 11:1297–1322.

Selten, R. 1998. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics* 1:43–61.

Shaw, A. D.; Horton, J. J.; and Chen, D. L. 2011. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*, 275–284.

Witkowski, J., and Parkes, D. 2011. Peer Prediction with Private Beliefs. In *Proceedings of the 1st Workshop on Social Computing and User Generated Content (SC'11)*.