A Robust Bayesian Truth Serum for Small Populations [†]

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 mechanisms 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 number of agents $n \geq 2$, but relies on a *common prior*, shared by all agents and the mechanism. The *Bayesian Truth Serum* (BTS) relaxes this assumption. While BTS still assumes that agents share a common prior, this prior need not be known to the mechanism. However, BTS is only incentive compatible for a large enough number of agents, and the particular number of agents required is uncertain because it depends on this private prior. In this paper, we present a *robust BTS* for the elicitation of binary information which is incentive compatible for every $n \geq 3$, taking advantage of a particularity of the quadratic scoring rule. The robust BTS is the first peer prediction mechanism to provide strict incentive compatibility for every $n \geq 3$ without relying on knowledge of the common prior. 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, 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 a user's 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. A significant drawback of BTS is that it only aligns incentives for a large enough number of agents, where 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 is not numerically robust for all inputs.

In this paper, we present the *robust Bayesian Truth Serum* (RBTS) mechanism, which, to the best of our knowledge, is the first peer prediction mechanism that does not rely on knowledge of the common prior to provide strict incentive compatibility for every 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. The main limitation of RBTS relative to earlier mechanisms, is that it applies only to the elicitation of *binary information*; e.g., good or bad experiences, or true or false classification labels.[1] Extending RBTS to incorporate more than two signals is the most important direction for future research.

[†]This version of the paper is different from the one published in the Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12) in that it provides an alternative proof for Lemma 5 and corrects typos in Example 3.

[1]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). Indeed, as the number of possible information reports increases, so does the difficulty imposed on users in providing the prediction report, which must include estimates for the additional possible information reports.

RBTS takes the same reports as BTS, and an agent's payment continues to consist of one component that depends on an agent's information report and a second component that depends on an agent's prediction report. The main innovation 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 particularity of the quadratic scoring rule, by which an agent prefers a shadow posterior 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 posterior belief and the agent's prediction report to the quadratic scoring rule, adopting the information report of a third agent k as the event to be predicted.

Related Work

In addition to the original peer prediction method and the original BTS, there is other related work. Jurca and Faltings (2007) extend the original peer prediction method to allow agents to have small deviations from a common prior while still assuming this prior is known to the mechanism, establishing a trade-off between the required budget and the robustness to 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 mechanism.

Jurca and Faltings (2008) assume a common prior known to the agents but unknown to the mechanism in an *online* polling setting, where the current empirical frequency of reports is published and updated as agents arrive.[2] While their mechanism only requires an information report (and not a prediction report), it is not incentive compatible. Rather, agents must behave strategically in deciding how to report information to the mechanism. Indeed, Jurca and Faltings (2011) give an impossibility result in regard to achieving incentive compatibility when the prior is not known to the mechanism and when agents make only an information 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, agents are indifferent between being truthful and misreporting. For this reason it does not extend to settings in which providing accurate information is costly or when agents have some other outside incentive for making false reports.

Witkowski and Parkes (2012) extend the methods of the present paper to a setting without a common prior to agents, obtaining incentive compatibility for every $n \geq 2$ and binary information. A crucial new requirement is one of "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.

[2]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).

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: *states* and *signals*. The state 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 on $\{0, 1\}$; sometimes the range is denoted $\{l, h\}$ 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 a *common prior*, consisting of $\Pr(T = t)$ and $\Pr(S = h \mid T = t)$, the conditional probability of observing a high signal given each possible state t. We require that the prior is admissible:

Definition 1. *The common prior is* admissible *if it satisfies the following properties:*

- *There are two or more possible states; i.e.,* $m \geq 2$.
- *Every state has positive probability, so that* $\Pr(T = t) > 0$ *for all* $t \in \{1, \ldots, m\}$.
- *States are distinct, such that* $\Pr(S = h \mid T = t) \neq \Pr(S = h \mid T = t')$ *for any two* $t \neq t'$. *We adopt the convention that states are sorted; i.e.,* $\Pr(S = h \mid T = 1) < \ldots < \Pr(S = h \mid T = m)$. *We refer to this as the assortative property.*
- *The signal beliefs conditional on state are fully mixed, with* $0 < \Pr(S = h \mid T = t) < 1$ *for all* t.

Admissibility is a weak requirement. In particular, note that any prior can be transformed into an admissible prior as (1) all signals beliefs conditional on state are fully mixed for states with positive probability, and (2) the signal beliefs conditional on state are distinct for at least two states with positive probability. Any two states with the same signal belief probability can be merged into a new state, and states with zero probability can be dropped. The mechanism does not need any knowledge about the prior beyond admissibility.

Given an agent i's realized signal s_i, the agent can update her posterior belief $\Pr(S_j = h \mid S_i = s_i)$ 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 $p_{\{h\}} = \Pr(S_j = h \mid S_i = h)$ and $p_{\{l\}} = Pr(S_j = h \mid S_i = l)$, respectively. We refer to this as a "first order" signal posterior, and have

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

where the posterior on state 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

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

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 example, for agent i 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 state T, $Pr(T = t|S_i = s_i)$, which becomes the belief 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).[3] While we present the binary version of this mechanism, BTS is defined for an arbitrary number of signals.

In BTS, 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 in BTS involves three steps:

1. For every agent $j \neq i$, calculate the arithmetic mean[4] of all agents' signal reports except those of agents i and j:

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

2. For every agent $j \neq i$, calculate the geometric mean of all prediction reports except those from i and j, on both high and low signals,

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

3. Calculate the BTS score for agent i:

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

[3] 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.

[4] Prelec adopts Laplacian smoothing to avoid zero values.

This simplifies for $n \to \infty$ in that the summations over $j \neq i$ in Equation 6 can be replaced with the information and prediction scores computed using just one, randomly selected, $j \neq i$.

The Bayesian Truth Serum mechanism is *strictly 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 strictly 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 need not align incentives for small groups of agents. Moreover, it need not satisfy *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 $n = 3$.*

This limitation of BTS can be understood from Prelec's treatment of BTS. Generally the number of agents required for BTS to be Bayes-Nash incentive compatible depends on the prior and is hard to characterize. Still, BTS has been discussed in various places without noting this important caveat, e.g., (Jurca and Faltings 2008; Chen and Pennock 2010). For this reason, we provide a concrete example. The example is not unique, and does not rely on $n = 3$.

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

Consider agent $i = 1$, 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 (6) that correspond to agent $j = 2$ when agent 1 reports truthfully is:

$$E\left[\ln(\frac{\bar{X}_{-12}}{\bar{Y}_{-12}}) + \bar{X}_{-12} \ln(\frac{0.764}{\bar{X}_{-12}}) + (1 - \bar{X}_{-12}) \ln(\frac{1 - 0.764}{1 - \bar{X}_{-12}})\right],$$

with the expectation taken with respect to random variables $\bar{X}_{-12}$ and $\bar{Y}_{-12}$. With probability $p_{\{h\}} = 0.764$, agent 1 believes that $\bar{x}_{-12} = (1 + 1)/3 = 2/3$ and $\bar{y}_{-12} = 0.764$ and with probability $1 - p_{\{h\}} = 0.236$ that $\bar{x}_{-12} = (0 + 1)/3 = 1/3$ and $\bar{y}_{-12} = p_{\{l\}} = 0.339$. Given this, 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 in (6), the total expected score when agent 1 is truthful is -0.450. BTS fails interim IR.

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 (6), yields a total expected score of 0.298. Agent 1 can do better by making a misreport.

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 variables $\overline{X}_{-ij}$, $\overline{Y}_{-ij}$ and $\bar{Y}'_{-ij}$ take on their respective values with probability 1:

$$\begin{aligned}
\overline{X}_{-1j} &= \lim_{n\to\infty} \frac{1}{n}\left((n-2)p_{\{h\}} + 1\right) = p_{\{h\}} \\
\overline{Y}_{-1j} &= \lim_{n\to\infty} \left((p_{\{h\}}^{(n-2)p_{\{h\}}})(p_{\{l\}}^{(n-2)(1-p_{\{h\}})})\right)^{1/(n-2)} \\
&= (p_{\{h\}}^{p_{\{h\}}})(p_{\{l\}}^{1-p_{\{h\}}}) = 0.631, \\
\bar{Y}'_{-1j} &= (1-p_{\{h\}})^{p_{\{h\}}}(1-p_{\{l\}})^{1-p_{\{h\}}} = 0.301.
\end{aligned}$$

If agent 1 reports truthfully ($x_1 = 1$ 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 (and thus total score) of $\ln(\frac{0.236}{0.301}) = -0.243$. BTS is Bayes-Nash incentive compatible in the large n limit in the example.

Having demonstrated the failure of incentive alignment and interim IR for small n 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 the Robust Bayesian Truth Serum (RBTS). RBTS is incentive compatible for every $n \geq 3$, *ex post* individually rational (meaning no agent's payment is negative, for any outcome), and numerically robust. We first introduce 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 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. *(e. g., 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 a *reference* agent $j = i+1$ (modulo n) and a *peer* agent $k = i+2$ (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 $\delta = \min(y_j, 1 - y_j)$. 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 same setting as in 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 agents 2 and 3 play the roles of reference j and peer k, respectively. We assume agents 2 and 3 are truthful. We first illustrate the calculations when $S_1 = h$, $S_2 = l$, and $S_3 = h$. If agent 1 is truthful, we have $y'_1 = y_2 + \delta = 0.339 + 0.339 = 0.678$ since $y_2 = 0.339$ and $\delta = 0.339$. Since $x_3 = 1$, agent 1's information score is $2y'_1 - {y'_1}^2 = 2(0.678) - 0.678^2 = 0.896$. Since $y_1 = 0.764$ and $x_3 = 1$, the prediction score is $2(0.764) - 0.764^2 = 0.944$. In total, the agent's score is 1.84.

To establish that, when $S_1 = h$, agent 1 is best off reporting truthfully, we need to consider the expected score and thus the distribution on 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. Agent 1's expected *prediction score* is $0.764(2(0.764) - 0.764^2) + 0.236(2(0.236) - 0.236^2) = 0.820$. 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 and

agent 1's expected information score is:

$$\begin{aligned}
&\Pr(S_2 = h \mid S_1 = h) \\
&\quad\big[\ \Pr(S_3 = h \mid S_1 = h, S_2 = h) R_q(0.764 + 0.236, 1) \\
&\qquad + \Pr(S_3 = l \mid S_1 = h, S_2 = h) R_q(0.764 + 0.236, 0)\big] \\
&+ \Pr(S_2 = l \mid S_1 = h) \\
&\quad\big[\ \Pr(S_3 = h \mid S_1 = h, S_2 = l) R_q(0.339 + 0.339, 1) \\
&\qquad + \Pr(S_3 = l \mid S_1 = h, S_2 = l) R_q(0.339 + 0.339, 0)\big] \\
&= \quad p_{\{h\}}\big[p_{\{h,h\}}\left(2(1) - 1^2\right) + \left(1 - p_{\{h,h\}}\right)\left(1 - 1^2\right)\big] \\
&\quad + \left(1 - p_{\{h\}}\right)\big[p_{\{h,l\}}\left(2(0.678) - 0.678^2\right) \\
&\qquad\qquad + \left(1 - p_{\{h,l\}}\right)\left(1 - 0.678^2\right)\big] = 0.79.
\end{aligned}$$

For a report of $x_1 = 0$, the expected information score is:

$$\begin{aligned}
&p_{\{h\}}\big[\qquad p_{\{h,h\}} R_q(0.764 - 0.236, 1) \\
&\qquad + \left(1 - p_{\{h,h\}}\right) R_q(0.764 - 0.236, 0)\big] \\
&+ \left(1 - p_{\{h\}}\right)\big[\qquad p_{\{h,l\}}\ R_q(0.339 - 0.339, 1) \\
&\qquad + \left(1 - p_{\{h,l\}}\right) R_q(0.339 - 0.339, 0)\big] = 0.664
\end{aligned}$$

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

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

Incentive Compatibility

In establishing the incentive compatibility of RBTS, we begin with some technical lemmas. The first lemma also establishes *stochastic relevance*, so that the signal posteriors are distinct for distinct signal observations. We then introduce a proper scoring rule for eliciting signals rather than belief reports, and use this as a building block for analyzing RBTS.

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

Proof. The fully mixed property of admissible priors ensures that beliefs are always interior, and $1 > p_{\{h\}} > 0$ and $1 > p_{\{l\}} > 0$. Furthermore, if $p_{\{h\}} > \Pr(S_j = h)$, then this implies $\Pr(S_j = h) > p_{\{l\}}$ since

$$\begin{aligned}
\Pr(S_j = h) = \quad & p_{\{h\}} \Pr(S_i = h) \\
& + \Pr(S_j = h | S_i = l) \Pr(S_i = l) \\
\Leftrightarrow \quad & \Pr(S_j = h)\left(1 - p_{\{h\}}\right) = p_{\{l\}}\left(1 - \Pr(S_j = h)\right) \\
\underbrace{\Rightarrow}_{p_{\{h\}} > \Pr(S_j = h)} \quad & \Pr(S_j = h) > p_{\{l\}}.
\end{aligned}$$

Left to show is $p_{\{h\}} > \Pr(S_j = h)$ given admissible priors. The remainder of the proof proceeds in three steps:

First, associate every state with one of two groups H and L. Associate states $t \in \{1, \ldots, m\}$ for which $\Pr(S = h|T = t) > \Pr(S = h)$ with group H, and states $t \in \{1, \ldots, m\}$ for which $\Pr(S = h|T = t) \leq \Pr(S = h)$ with group L. That is, the states in group H are those that put more weight on signal h than the signal prior, and the states in group L are those that put less or equal weight on signal h than the signal prior.

Second, verify that both H and L are non-empty, i.e. there is at least one state t and at least one other state $t' \neq t$ with $t, t' \in \{1, \ldots, m\}$, such that $\Pr(S = h|T = t) > \Pr(S = h)$ and $\Pr(S = h|T = t') \leq \Pr(S = h)$. To see this, first see that $\Pr(S = h|T = t) = \Pr(S = h)$ for all t is excluded by admissiblity which requires that states are distinct, i.e. have different signal conditionals. Then, recall that the signal prior $\Pr(S = h)$ is the (weighted) *average* of signal conditionals $\Pr(S = h|T = t)$ (Equation 3):

$$\Pr(S = h) = \sum_{t=1}^{m} \Pr(S = h|T = t) \cdot \Pr(T = t).$$

This excludes that $\Pr(S = h|T = t) > \Pr(S = h)$ for all $t \in \{1, \ldots, m\}$, that $\Pr(S = h|T = t) < \Pr(S = h)$ for all $t \in \{1, \ldots, m\}$, or that $\Pr(S = h|T = t') = \Pr(S = h)$ for *some* $t' \in \{1, \ldots, m\}$ and $\Pr(S = h|T = t) < \Pr(S = h)$ for all other $t \neq t'$. Moreover, verify that both groups, L and H, have positive probability since admissibility demands that every state has positive probability, i.e. $\Pr(T = t) > 0$ for all $t \in \{1, \ldots, m\}$.

Third, the probability of states in group H increases given observation $S = h$. To see this, we need to know for which $t \in \{1, \ldots, m\}$ is $\Pr(T = t|S = h) > \Pr(T = t)$ and obtain:

$$\begin{aligned}
&\Pr(T = t|S = h) > \Pr(T = t) \\
\Leftrightarrow\ &\frac{\Pr(S = h|T = t) \cdot \Pr(T = t)}{\Pr(S = h)} > \Pr(T = t) \\
\Leftrightarrow\ &\Pr(S = h|T = t) > \Pr(S = h).
\end{aligned}$$

That is, exactly those states in group H become more likely after signal h. The statement $p_{\{h\}} > \Pr(S_j = h)$ follows because states in group H have more weight on signal h than the signal prior $\Pr(S = h)$ and become more likely after $S = h$.

□

We extend this observation 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 state posterior $\Pr(T = t \mid S_i = h)$ taking the role of state prior, $\Pr(T = t)$, in the analysis. Since $p_{\{h,l\}} = p_{\{l,h\}}$, the other case $p_{\{l,h\}} > p_{\{l\}} = p_{\{l,l\}}$ can be shown analogously. □

Lemma 7 is a known result for which we present a proof only to build intuition.

Lemma 7. *(e. g., Selten, 1998) 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* $|y - p|$*.*

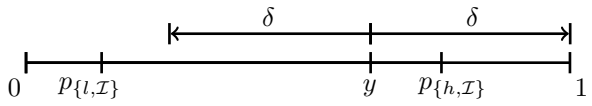

Figure 1: Illustration of the shadowing method with $y \in (p_{\{l,\mathcal{I}\}}, p_{\{h,\mathcal{I}\}})$. Note that $p_{\{l,\mathcal{I}\}}$ is closer to $y'_i = y - \delta$ than to $y'_i = y + \delta$, and that $p_{\{h,\mathcal{I}\}}$ is closer to $y'_i = y + \delta$ than to $y'_i = y - \delta$.

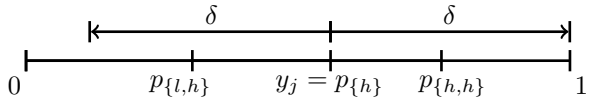

Figure 2: An illustration of RBTS in the $S_j = h$ case. Note that y_j is always strictly in between agent i's two possible second-order posteriors $p_{\{l,h\}}$ and $p_{\{h,h\}}$.

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. The logarithmic rule, for example, does not satisfy this property.

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

Making a step towards our main result, we adapt proper scoring rules to elicit *signals* (rather than *beliefs*) truthfully.

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 a general form of the method, we allow agent i to have observed a sequence of signals $\mathcal{I} \in \{0, 1\}^o$ (for some $o \in \{0, 1, \ldots\}$) before observing a new signal S_i. The *shadowing method* then proceeds as:

1. Agent i receives a signal $S_i \in \{0, 1\} = \{l, h\}$ and, based on the prior and previously-observed signals $\mathcal{I}$, forms a posterior belief $p \in \{p_{\{l,\mathcal{I}\}}, p_{\{h,\mathcal{I}\}}\}$ about ω.
2. The center asks the agent for signal report $x_i \in \{0, 1\}$ and transforms it into a "shadow" posterior report y'_i by:
$$y'_i = \begin{cases} y + \delta, & \text{if} \quad x_i = 1 \\ y - \delta, & \text{if} \quad x_i = 0, \end{cases} \tag{9}$$
where $y \in (0, 1)$ is a parameter of the method and $\delta = \min(y, 1 - y)$ (also see Figure 1).
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:
$$R_q(y'_i, \omega). \tag{10}$$

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. The proof is via reasoning about the distance between the agent's posterior and the respective shadow posterior. 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 strictly Bayes-Nash incentive compatible for any $n \geq 3$ and all admissible priors.*

Proof. Fix some i, reference j and peer k, and assume agents j and k report truthfully. It needs to be shown that it is the unique best response for agent i to report truthfully. The best response conditions for x_i and y_i can be analyzed for each report type separately, because y_i affects only the prediction score, and x_i affects only the information score. Noting that strict incentives for the prediction report y_i follow directly from the use of the strictly proper quadratic scoring rule, we focus on x_i. There are two cases to consider in regard to agent j:

1. $S_j = h$ and so $y_j = p_{\{h\}}$ (also see Figure 2). 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 $y_j = p_{\{l\}}$. 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. □

Other Properties and Discussion

Theorem 10. *The scores in the Robust Bayesian Truth Serum are in $[0, 2]$ for any reports from agents including any $y_i \in [0, 1]$, and thus RBTS is ex post individually rational and numerically robust.*

Proof. The binary quadratic scoring rule $R_q(y, \omega)$ is well-defined for any input $y \in [0, 1]$ and $\omega \in \{0, 1\}$, and generates scores on $[0, 1]$. The inputs to R_q for computing the information score are $y := y'_i \in [0, 1]$ and $\omega := x_k \in \{0, 1\}$. The inputs for computing the prediction 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. This is the case where BTS is not well defined.

For a designer with a particular budget $B > 0$, a straightforward extension of RBTS 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. This extension to RBTS remains strictly incentive compatible when the agents do not know which of them is the excluded agent. While multiple equilibria cannot be avoided in peer prediction settings without trusted reports, this randomized extension ensures that the agents' scores in the truthful equilibrium cannot be less than in any other equilibrium. Moreover, by sacrificing *ex post* individual rationality, the same technique can be used to implement a mechanism with $B = 0$.

In contrast to BTS, RBTS easily adapts to *online* polling settings, where the center publishes partial information about reports 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 every 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 RBTS. 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). Most important for future work is to relax the requirement of binary information reports, which is the limitation of RBTS in comparison to BTS.

Acknowledgments

We thank Yiling Chen for early discussions on the Bayesian Truth Serum, and the anonymous reviewers for useful comments and feedback. This work is supported in part by NSF Grant No. CCF-0915016 and a Microsoft Faculty Grant. Jens Witkowski is also grateful for financial support through a PhD fellowship from the Landesgraduiertenförderung Baden-Württemberg.

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 (JMLR)* 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. 2012. Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*.