Minimum Truth Serums with Optional Predictions

BLAKE RILEY, University of Illinois at Urbana-Champaign

In this paper, I introduce a class of mechanisms for eliciting private correlated signals from a group of expected score maximizers without external verification or knowledge about the agents' belief structure. Built on proper scoring rules, these *minimum truth serums* ask agents to report a signal and a prediction of the signals of others. If two agents with the same signal have the same expectations about the signals of others, the Bayesian incentive compatibility of these mechanisms follows with no further assumptions on the agents' belief structure. With a slight modification, the mechanism is still feasible and incentive compatible when the prediction portion of the report is optional.

General Terms: Theory, Economics

Additional Key Words and Phrases: Information elicitation, mechanism design, Bayesian truth serum, peer prediction, proper scoring rules

Author's address: blake.j.riley@gmail.com

1. INTRODUCTION

The Bayesian truth serum, introduced by Prelec [2004], was one of the first mechanisms for eliciting private information from a group of agents regarding subjective or hypothetical questions without external verification. The mechanism operates by asking agents for an *information report* from a finite set of answers—corresponding to the agent's private signal—and a *prediction report* of the distribution of signals of other agents—corresponding to the agent's first-order posterior beliefs—and assigning scores to agents based on the collective reports. The original truth serum has many desirable properties, such as being

- *detail free*, requiring zero knowledge of the agents' common prior to determine scores,
- *interim individually rational*, giving each agent a non-negative expected payment conditional on their private information, and
- *collusion resistant*, with the truth-telling equilibrium being interim Pareto-dominant among all Bayes-Nash equilibria.

The mechanism can also be *ex-post budget balanced*, with total payments summing to zero for every possible realization of reports, though at the cost of individual rationality.

Many potential concerns about the robustness of the Bayesian truth serum remain. First, incentive compatibility holds only for a sufficiently large population, with the necessary size depending on the agents' unknown prior. Second, the mechanism could require arbitrarily large payments from the agents. Third, the assumption of common priors might be overly strong. Finally, requiring a prediction report in addition to an information report could make the mechanism too complex and unwieldy for some agents. Since Prelec's original work, multiple papers have sought to alleviate these concerns.

Contemporaneously with Prelec, the *peer-prediction mechanism* of Miller et al. [2005] depends only on an information report but is not detail-free, assuming precise knowledge of the posterior beliefs of agents for each signal. Jurca and Faltings [2011] investigate detail-free mechanisms that depend only on an information report, showing incentive compatibility is not possible in general, and develop instead a notion of *helpful reporting*. Jurca and Faltings [2007] and Carvalho and Larson [2012] discuss various forms of collusion resistance in peer-prediction mechanism.

Addressing Prelec directly, Witkowski and Parkes [2012a] introduce the *robust Bayesian truth serum* for the special case of binary signals, which is incentive compatible for $n \geq 3$ agents and has bounded payments. Under the additional assumption that agents' beliefs could be elicited both before and after receipt of their private signal[1], Witkowski and Parkes [2012b] provide a mechanism that is incentive compatible for $n \geq 2$ agents while eliminating the common prior assumption, again for binary signals. In both papers, Witkowski and Parkes favor ex-post individual rationality over ex-post budget-balance, a reasonable choice since the two properties are incompatible in all non-trivial mechanisms for this setting.

The approaches of Radanovic and Faltings [2013] and Zhang and Chen [2014] are the most similar to this paper. Each introduce mechanisms that are incentive compatible for small groups agents for any finite number of signals, extending the binary truth serum of Witkowski and Parkes [2012a]. Furthermore, they consider correlated signals in general, not conditionally independent signals as in Prelec or Witkowski and Parkes.

Radanovic and Faltings prove no mechanism can be incentive compatible for all belief structures if it uses only information reports or is *decomposable* (i.e. additively separable in the information and prediction reports). They give sufficient conditions for each respective type of mechanism to be incentive compatible, which can be roughly interpreted as each signal being sufficiently positively correlated with itself across agents. Radanovic and Faltings then introduce mechanisms that meet these necessary conditions for $n \geq 2$ agents.

The mechanism of Zhang and Chen on the other hand puts no constraints on the correlation between signals, requiring instead a second-order stochastic relevance condition in addition to common priors for $n \geq 3$ agents. They achieve this by operating the mechanism sequentially, first collecting signals and then collecting predictions after passing one signal report to each agent.

In this paper, I introduce a class of non-decomposable truth serums where assuming *common predictions* is sufficient for weak incentive compatibility. The common predictions assumption is similar in spirit to common priors, but is neither weaker nor stronger. Under additional mild assumptions—stochastic relevance of signals, full support of agent beliefs, and the use of a strictly proper scoring rule—the mechanism is strictly incentive compatible when the number of agents is one more than the number of possible signals. I make no assumptions about the degree or direction of correlation between signals. Furthermore, with minor modifications, the mechanism is still feasible if prediction reports are optional. Agents will have a strict incentive to make a prediction even if others might omit theirs. I also address some potential concerns about the robustness of the mechanism, like how to eliminate Pareto-dominating uninformative equilibria and how to weaken the assumption of risk neutrality.

2. SETTING

The respondent pool contains n rational, risk-neutral, and self-interested agents. Let a typical agent have index i. Each agent receives a *signal* x_i drawn from a shared finite set T. The vector $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_n) \in T^n$ is the *signal profile* of the participants. Signals are private, with each agent knowing their own with certainty and having probabilistic beliefs about the signals of others. Signals can represent an attribute of the agent or observations about an external state of the world, depending on the question of interest to the mechanism operator. For instance, college freshmen could be asked to report the number of drinks of alcohol they've had in the past week from the set $T = \{$*0 drinks, 1-2 drinks, 3-6 drinks, 7+ drinks*$\}$. Alternatively, crowdsourced reviewers could be asked to evaluate a writing sample according to some task guidlines, reporting a rating from the set $T = \{$*1, 2, 3, 4, 5*$\}$. In this last example, a signal of $x_i =$ *4* reflects i's judgement that the writing sample fits into the *4* category given their

[1] For instance, before and after receiving an item purchased online.

observations, although i could still be uncertain about what rating the writing "truly" deserves.

Signals are correlated across agents. Conditional on their private signal, each agent has a posterior *prediction* $p_i = p(x_i) = \mathrm{E}_i[\bar{x}_{\text{-}i} \,|\, x_i] \in \Delta^T$ of the distribution of others' signals, where $\bar{x}^t_{\text{-}i} = \sum_{j \neq i} \mathbb{I}(x_j = t)/(n-1)$ is the sample proportion of the agents except for i receiving signal t and $\bar{x}_{\text{-}i} = (\ldots, \bar{x}^t_{\text{-}i}, \ldots) \in \Delta^T$ is the vector of sample proportions. I will use $p_i = p(x_i)$ to refer to i's actual prediction and $p(\hat{x}_i)$ to refer to i's prediction conditional on a hypothetical signal $\hat{x}_i$. The vector $\boldsymbol{p} = (p_1, \ldots, p_n) \in (\Delta^T)^n$ is the *prediction profile* of agents. Note that I consider any correlated signals similarly to Zhang and Chen [2014], generalizing the conditionally-independent signals of Prelec or Witkowski and Parkes.

To be meaningful, different signals should convey some distinct information, as expressed in the following assumption:

Definition 2.1 (*Stochastic relevance*). Signals are *stochastically relevant* if different signals induce different predictions:

$$\forall i, j: \quad x_i \neq x_j \quad \Longrightarrow \quad p(x_i) \neq p(x_j)$$

Furthermore, private signals are the only source of differences in beliefs:

Definition 2.2 (*Common predictions*). Agents have *common predictions* if agents with the same signal have the same posterior predictions:

$$\forall i, j: \quad x_i = x_j \quad \Longrightarrow \quad p(x_i) = p(x_j)$$

Common predictions is the converse of stochastic relevance, and together they imply a bijection between signals and predictions across agents.

For incentives to be strict, agents should think every profile of signals is possible:

Definition 2.3 (*Full support*). Agents' beliefs have *full support* if

$$\forall i,\, \forall(\ldots, t_{i-1}, t_{i+1}, \ldots) \in T^{n-1}: \quad \mathrm{Pr}_i[x_{\text{-}i} = (\ldots, t_{i-1}, t_{i+1}, \ldots) \,|\, x_i] > 0$$

The variable $\delta_i = \min_{t \in T} |\{x_j = t \,|\, j \neq i\}|$ denotes the minimum number of agents except for i in each signal group. The agents will frequently condition on at least one other agent reporting each signal, so let $p_{i|\delta_i \geq 1} = \mathrm{E}[\bar{x}_{\text{-}i} \,|\, x_i \textit{ and } \delta_i \geq 1]$ and $p_{i|\delta_i = 0} = \mathrm{E}[\bar{x}_{\text{-}i} \,|\, x_i \textit{ and } \delta_i = 0]$. Then, the agent's prediction p_i satisfies

$$p_i = \mathrm{Pr}_i[\delta_i = 0 \,|\, x_i]\; p_{i|\delta_i=0} + \mathrm{Pr}_i[\delta_i \geq 1 \,|\, x_i]\; p_{i|\delta_i \geq 1}$$

The mechanism collects signals and predictions from all agents simultaneously, assigning agents a score $S_i(\boldsymbol{x}, \boldsymbol{p})$ based on the vectors of reports $\boldsymbol{x}$ and $\boldsymbol{p}$. Agents are risk-neutral and maximize their expected score from the mechanism conditional on their private information. The score can be interpreted as a payment from the mechanism to the agent—with negative scores being payments from the agent to the mechanism—or an abstract reputation score.

2.1. Common Predictions vs Common Priors

The common predictions assumption is distinct from the common prior assumption used extensively in this literature. The two are still similar in spirit, both saying that differences in beliefs come only from differences in observations. Although I will not assume common priors in this paper, I define it here for comparison:

Definition 2.4 (*Common priors*). Agents have *common priors* if all assign the same prior probability to each signal profile:

$$\forall i, j, \forall(t_1, \ldots, t_n) \in T^n: \quad \mathrm{Pr}_i[\boldsymbol{x} = (t_1, \ldots, t_n)] = \mathrm{Pr}_j[\boldsymbol{x} = (t_1, \ldots, t_n)]$$

Common priors is neither necessary nor sufficient for common predictions. A common prior where agents have different marginal distributions will not satisfy common predictions in general, though it could if some agents have zero probability of receiving some signals due to the ambiguity of conditioning on a zero probability event. Agents can have common predictions and uncommon priors if an agent disagrees with others about the prior marginal distribution of his own signal.

When agents have common predictions and common priors, beliefs will have some degree of symmetry across agents, though an exact characterization is beyond the scope of this paper. An *exchangeable* common prior—with $\Pr_i[\boldsymbol{x}] = \Pr_i[\sigma(\boldsymbol{x})]$ for every signal profile $\boldsymbol{x}$ and every permutation $\sigma(\cdot)$—is clearly sufficient for common predictions. For a simple example of a non-exchangeable common prior with common predictions, consider three agents with $T = \{a, b, c\}$ where a profile with three identical signals has probability $15/102$, three profiles satisfy $\Pr[(a, b, c)] = \Pr[(c, a, b)] = \Pr[(b, c, a)] = 12/102$, and the remaining 21 profiles have probability $1/102$ for each. An agent receiving $x_i = a$ has the prediction $p(a) = (1/2, 1/4, 1/4)$. However, $\Pr[x_2 = b \,|\, x_1 = a] = 42/102 \neq 9/102 = \Pr[x_3 = b \,|\, x_1 = a]$, so agents 2 and 3 are not identical from agent 1's perspective.

3. PROPER SCORING RULES

Truth serums have their foundations in scoring rules. Proper scoring rules are incentive schemes for eliciting a rational, risk-neutral payment-maximizer's honest probabilities of some event. For an overview of the theory of scoring rules and their applications, see [Gneiting and Raftery 2007]. The event in question is usually assumed to be publicly observed or externally verified, such as the weather tomorrow or the winner of an election. In this paper, however, the agents will be scored on their predictions of the reported signals of other agents.

A scoring rule R assigns payments $R(t, \hat{p}_i)$ for a realized outcome t of a random variable x and a reported probability distribution $\hat{p}_i$. If the agent's expected score $E[R(x, \hat{p}_i)]$ is maximized when they report their honest subjective probabilities of each event, the scoring rule is *proper*. Scoring rules are *strictly proper* if the honest prediction is the unique maximizer of the expected score and *weakly proper* if other reports can also be maximizers.

Well-known examples of strictly proper scoring rules for discrete random variables include the logarithmic rule [Good 1952]

$$R_{\text{log}}(t, p_i) = \ln(p_i^t)$$

and the quadratic scoring rule [Brier 1950]

$$R_{\text{quad}}(t, p_i) = 1 - \frac{1}{2}\sum_{s \in T}(\mathbb{I}(t = s) - p_i^s)^2$$

where p_i^t is the probability assigned to realization t.

If R_1 and R_2 are proper scoring rules, then the affine combination

$$R(t, p_i) = a_1 R_1(t, p_i) + a_2 R_2(t, p_i) + b$$

is also a proper scoring rule for all $a_1, a_2 \geq 0$ and $b \in \mathbb{R}$. Using this property, a scoring rule for the probability of a multinomial event can be easily extended to a scoring rule for the expected proportion of outcomes by averaging over scores for each individual event. For instance, the extended logarithmic scoring rule is

$$R_{\text{log}}(\bar{x}, p_i) = \sum_{t \in T} \bar{x}^t \ln(p_i^t)$$

and the extended quadratic scoring rule is

$$R_{\text{quad}}(\bar{x}, p_i) = \frac{1}{2} + \sum_{t \in T} p_i^t \bar{x}^t - \frac{1}{2}(p_i^t)^2$$

where $\bar{x}^t$ is the proportion of t outcomes in the sample. Extended scoring rules are affine in their first argument and are proper when

$$\forall p_i, \hat{p}_i \in \Delta^T, \quad R(p_i, p_i) \geq R(p_i, \hat{p}_i)$$

All extended proper scoring rules for discrete events can be represented using some convex function $F : \Delta^T \to \mathbb{R}$ [Savage 1971] as

$$R(\bar{x}, p_i) = F(p_i) + \langle F'(p_i), \bar{x} - p_i \rangle$$

where $F'(p)$ is a subgradient of F (supporting F from below analogous to the gradient of a differentiable function) and $\langle \cdot, \cdot \rangle$ is the standard inner product. The function $F(p)$ can be interpreted as the score an agent with belief p expects to receive when reporting honestly. For instance, we have

$$R_{\text{log}}(\bar{x}, p) = \sum_{t \in T} p^t \ln(p^t) + \left\langle (\ldots, \ln(p^t) + 1, \ldots), \bar{x} - p \right\rangle, \text{ for } F(p) = \sum_{t \in T} p^t \ln(p^t) \text{ and}$$

$$R_{\text{quad}}(\bar{x}, p) = \frac{1}{2} + \frac{1}{2} \sum_{t \in T} (p^t)^2 + \langle p, \bar{x} - p \rangle, \text{ for } F(p) = \frac{1}{2} + \frac{1}{2} \sum_{t \in T} (p^t)^2$$

4. MINIMUM TRUTH SERUMS

A truth serum is a detail-free mechanism for collecting private signals from a group of agents without external verification. The mechanism operates by soliciting from each agent i their signal x_i and prediction p_i of the distribution of signals of other agents. Given some proper scoring rule R, the agents' predictions will be scored as $R(\bar{x}_{\text{-}i}, p_i)$ based on the actual distribution of other signals $\bar{x}_{\text{-}i}$. Since R is proper, this score will be maximized in expectation when the agent reports p_i honestly. For the agent to honestly reveal their signal x_i as well, the final score will be bounded above by the score for the "average" prediction of others reporting the same signal. The predictions of others with the same signal will be aggregated so that if all the inputs are identical, then that same value is the output, as expressed in the following definition:

Definition 4.1 (*Unanimous aggregator*). A function $g : \cup_{k \in \mathbb{N}} (\Delta^T)^k \to \Delta^T$ that maps profiles of predictions back into predictions is a *unanimous* or *idempotent aggregator* if $g(\{p_j, \ldots, p_j\}) = p_j$ for all $p_j \in \Delta^T$.

Common averaging functions such as the arithmetic mean or the normalized geometric mean are unanimous. Alternatively, the aggregator could select one input according to some criterion, such as an element in the set with minimum Euclidean norm breaking ties lexicographically.

The precise definition of the minimum truth serum class is as follows:

Definition 4.2 (*Class of minimum truth serums*). Given a proper scoring rule R and a unanimous aggregator g, a minimum truth serum is defined by the following procedure:

(1) Collect the vectors of signals $\boldsymbol{x}$ and predictions $\boldsymbol{p}$ from agents simultaneously, with elements x_i and p_i being the reports of agent i.
(2) Define the variables

$$\delta_i = \min_{t \in T} \, |\{x_j = t \,|\, j \neq i\}|$$

for each agent denoting the minimum number of other agents that can be found in each signal group.

(3) If $\delta_i \geq 1$, compute the proxy prediction

$$q_i(x_i) = g(\{p_j \in p_{-i} \mid x_j = x_i\})$$

(4) Assign scores to each agent as

$$S_i(\boldsymbol{x}, \boldsymbol{p}) = \begin{cases} R(\bar{x}_{-i}, p_i) & \text{if } \delta_i = 0 \\ \min\{R(\bar{x}_{-i}, p_i), R(\bar{x}_{-i}, q_i(x_i))\} & \text{if } \delta_i \geq 1 \end{cases}$$

When there are no other agents besides i reporting x_i, the proxy prediction clearly isn't well defined. However, due to the definition of δ_i, the mechanism will sometimes avoid comparing i to the others with the same signal even when proxy prediction can be computed. This is so that i cannot purposefully report a rare signal to avoid the proxy upper bound. Since δ_i depends only on the reports of others, i cannot directly manipulate it.

Definition 4.3 (*Bayesian incentive compatibility*). A truth serum is *Bayesian incentive compatible* if, for all i,

$$\mathrm{E}[S_i((x_i, x_{-i}), (p_i, p_{-i})) \mid x_i, p_i] \geq \mathrm{E}[S_i((\hat{x}_i, x_{-i}), (\hat{p}_i, p_{-i})) \mid x_i, p_i]$$

for all $x_i, \hat{x}_i, p_i, \hat{p}_i$ and is strictly incentive compatible if the inequality is always strict when $x_i \neq \hat{x}_i$ or $p_i \neq \hat{p}_i$.

THEOREM 4.4. *All minimum truth serums are Bayesian incentive compatible if agents have common predictions. If $n \geq |T| + 1$, R is strictly proper, beliefs have full support, and signals are stochastically relevant, then the truth serum is strictly incentive compatible.*

PROOF. Suppose all agents except i report their signals and predictions honestly. Then, the expected score of agent i when giving report $(\hat{x}_i, \hat{p}_i)$ (with all probabilities conditional on x_i) satisfies

$$\begin{aligned} \mathrm{E}[S_i((\hat{x}_i, x_{-i}), (\hat{p}_i, p_{-i}))] &= \mathrm{Pr}_i[\delta_i = 0] \sum_{x_{-i}:\delta_i=0} R(\bar{x}_{-i}, \hat{p}_i) \,\mathrm{Pr}_i[x_{-i} \mid \delta_i = 0] \\ &\quad + \mathrm{Pr}_i[\delta_i \geq 1] \sum_{x_{-i}:\, \delta_i \geq 1} \min\{R(\bar{x}_{-i}, p_i), R(\bar{x}_{-i}, q_i(\hat{x}_i)\} \,\mathrm{Pr}_i[x_{-i} \mid \delta_i \geq 1] \\ &\leq \mathrm{Pr}_i[\delta_i = 0] \sum_{x_{-i}:\, \delta_i=0} R(\bar{x}_{-i}, \hat{p}_i) \,\mathrm{Pr}_i[x_{-i} \mid \delta_i = 0] \\ &\quad + \mathrm{Pr}_i[\delta_i \geq 1] \sum_{x_{-i}:\, \delta_i \geq 1} R(\bar{x}_{-i}, \hat{p}_i) \,\mathrm{Pr}_i[x_{-i} \mid \delta_i \geq 1] \\ &= \sum_{x_{-i}} R(\bar{x}_{-i}, \hat{p}_i) \,\mathrm{Pr}_i[x_{-i}] = R(p_i, \hat{p}_i) \leq R(p_i, p_i) \end{aligned}$$

where the last line follows from p_i being i's expectation of $\bar{x}_{-i}$ and R being affine and maximized at $\hat{p}_i = p_i$ as a proper scoring rule. Since $p_j = p_i$ when $x_j = x_i$ by common predictions, we have $q_i(x_i) = g(\{p_j \in p_{-i} \mid x_j = x_i\}) = p_i$ because g is unanimous. Hence, the expected score when reporting truthfully achieves the upper bound above, and the honest report is a best response for i. Therefore, the minimum truth serum is (weakly) Bayesian incentive compatible. Notice that if $n < |T| + 1$, then δ_i is always zero and the agent will be indifferent between all information reports since there aren't enough other agents to fill each category.

Now assume $n \geq |T|+1$, R is strictly proper, beliefs have full support, and signals are stochastically relevant. Beliefs having full support and $n \geq |T|+1$ imply $\Pr_i[\delta_i \geq 1] > 0$, meaning there is some chance agent i will face the proxy upper bound. There are two cases to consider besides the honest report. First, any report with $\hat{p}_i \neq p_i$ will result in a strictly lower score since R is strictly proper. Second, reporting a dishonest signal $\hat{x}_i \neq x_i$ and an honest prediction p_i will lead to a strictly lower score since i will occasionally be matched with $q_i(\hat{x}_i) = p_j \neq p_i$ (by stochastic relevance), and we must occasionally have $R(\bar{x}_{\text{-}i}, q_i(\hat{x}_i)) < R(\bar{x}_{\text{-}i}, p_i)$ since R is strictly proper. Therefore, honesty is the unique best response under these assumptions. □

Although I've considered the number of agents n to be fixed, this could be a random variable from i's perspective. In this case, the assumption that $n \geq |T|+1$ can be replaced with $P_i[n \geq |T|+1] > 0$ for strict incentive compatiblity.

This mechanism is one of two known truth serums that can be used when a signal is negatively correlated with itself across agents[2]. Negative correlation can occur if there are a limited number of certain signals, where an observation by one agent "blocks" another agent's observation.

Consider the following example: A group of eleven birdwatching enthusiasts will attempt to sight the Lesser Jubjub, with each stationing themselves in a different area of a valley. Since the enthusiasts are prone to boasting in the absence of incentives for honesty, the president of their society will ask each whether they saw the bird (corresponding to a set of answers $T = \{\textit{Yes}, \textit{No}\}$) and give payments according to the minimum truth serum using the log scoring rule and some aggregator. Since the Lesser Jubjub is known to maintain a very small territory, if one watcher catches sight of it, she'll conclude it's less likely that others have seen it. In particular, assume beliefs are:

	p_i^{Yes}	p_i^{No}
$x_i = $ *Yes*	.05	.95
$x_i = $ *No*	.10	.90

Suppose the honestly reported signals are $x_1 = \textit{Yes}$ and $x_i = \textit{No}$ for all other i. Then, $\delta_1 = 0$ since all other ten reported *No*, and $\delta_i = 1$ for $i > 1$ since each answer was given by someone besides i. Finally, the score of agent 1 is $S_1(x,p) = R(\bar{x}_{-1}, p_1) = R((0.0, 1.0), (0.05, 0.95)) = \ln(0.95) \simeq -0.051$, and the scores of all other agents are $S_i(x,p) = \min\{R(\bar{x}_{\text{-}i}, p_i), R(\bar{x}_{\text{-}i}, g(x_i))\} = \min\{R((0.1, 0.9), (0.1, 0.9)), R((0.1, 0.9), (0.1, 0.9))\} = 0.1\ln(0.1) + 0.9\ln(0.9) \simeq -0.325$.

5. OPTIONAL PREDICTION REPORTS

Asking agents to provide an information report is straightforward since everyone has experience with surveys and ratings. However, requiring all agents to submit a prediction report could be a practical barrier to deploying a truth serum, especially for a moderately large set of answers. Impossibility results have been given [Radanovic and Faltings 2013] about mechanisms that depend only on information reports, which I skirt by including predictions, but making them optional. Because the minimum truth serum allows for the possibility of an agent being the only one in their signal group, the mechanism can be easily modified to operate when some predictions are missing. Rather than defining scores conditional on all signals being reported by some other agent, scores will be conditional on all signals being given by an agent who also made a prediction.

[2] The other being the knowledge-free peer prediction mechanism of Zhang and Chen [2014].

Definition 5.1 (*Minimum truth serum with optional predictions*). Given a bounded proper scoring rule R and a unanimous aggregator g, a minimum truth serum with optional predictions is defined by the following procedure:

(1) Collect the vectors of signals $\boldsymbol{x}$ and predictions $\boldsymbol{p}$ from agents simultaneously, where each agent has the option of selecting $p_i = \varnothing$.

(2) Define the variables

$$\delta_i = \min_{t \in T} \; |\{x_j = t \,|\, j \neq i \textit{ and } p_j \neq \varnothing\}|$$

for each agent denoting the minimum number of other agents with predictions that can be found in each signal group.

(3) If $\delta_i \geq 1$, compute the proxy prediction

$$q_i(x_i) = g(\{p_j \in p_{\text{-}i} \,|\, x_j = x_i \textit{ and } p_j \neq \varnothing\})$$

(4) Assign scores to each agent with $p_i \neq \varnothing$ as

$$S_i(\boldsymbol{x}, \boldsymbol{p}) = \begin{cases} R(\bar{x}_{\text{-}i}, p_i) & \text{if } \delta_i = 0 \\ \min\{R(\bar{x}_{\text{-}i}, p_i), R(\bar{x}_{\text{-}i}, q_i(x_i))\} & \text{if } \delta_i \geq 1 \end{cases}$$

(5) Assign scores to each agent with $p_i = \varnothing$ as

$$S_i(\boldsymbol{x}, \boldsymbol{p}) = \begin{cases} \min_{q \in \Delta^T}\{R(\bar{x}_{\text{-}i}, q)\} & \text{if } \delta_i = 0 \\ R(\bar{x}_{\text{-}i}, q_i(x_i)) & \text{if } \delta_i \geq 1 \end{cases}$$

giving the agent the minimum possible score according to R when $\delta_i = 0$.

THEOREM 5.2. *All minimum truth serums with optional predictions are Bayesian incentive compatible if agents have common predictions. If $n \geq |T| + 1$, R is strictly proper, beliefs have full support, and signals are stochastically relevant, then the truth serum is strictly incentive compatible.*

PROOF. In addition to the argument of the previous proof, we need to establish that each agent prefers giving a prediction to reporting $\hat{p}_i = \varnothing$. If all other agents report truthfully—but possibly with some giving the null prediction—the expected score for reporting $(\hat{x}_i, \varnothing)$ conditional on x_i is

$$\begin{aligned} \mathrm{E}[S_i((\hat{x}_i, x_{\text{-}i}), (\varnothing, p_{\text{-}i}))] &= \Pr_i[\delta_i = 0] \sum_{x_{-i} : \delta_i = 0} \min_{q \in \Delta^T}\{R(\bar{x}_{\text{-}i}, q)\} \Pr_i[x_{\text{-}i} \,|\, \delta_i = 0] \\ &\quad + \Pr_i[\delta_i \geq 1] \sum_{x_{-i} :\, \delta_i \geq 1} R(\bar{x}_{\text{-}i}, p(\hat{x}_i)) \Pr_i[x_{\text{-}i} \,|\, \delta_i \geq 1] \\ &\leq \Pr_i[\delta_i = 0] \sum_{x_{-i} :\, \delta_i = 0} R(\bar{x}_{\text{-}i}, p(\hat{x}_i)) \Pr_i[x_{\text{-}i} \,|\, \delta_i = 0] \\ &\quad + \Pr_i[\delta_i \geq 1] \sum_{x_{-i} :\, \delta_i \geq 1} R(\bar{x}_{\text{-}i}, p(\hat{x}_i)) \Pr_i[x_{\text{-}i} \,|\, \delta_i \geq 1] \\ &= \sum_{x_{-i}} R(\bar{x}_{\text{-}i}, p(\hat{x}_i)) \Pr_i[x_{\text{-}i}] = R(p_i, p(\hat{x}_i)) \leq R(p_i, p_i) = \mathrm{E}\, S_i(x_i, p_i) \end{aligned}$$

since $q_i(\hat{x}_i) = p(\hat{x}_i)$ by common predictions, so the agent is never better off when omitting the prediction. When beliefs have full support, the first inequality will be strict and providing a full, honest report is a strict best response. □

Notice that the proof concludes something slightly stronger than Bayesian incentive compatibility, showing a full and honest report is an interim best response even if others fail to best respond and omit their prediction.

While making a full report is preferable, the point of optional predictions is that we expect agents to occasionally omit them in practice. Let's consider the incentives of an agent conditional on reporting $\hat{p}_i = \varnothing$. Ideally, such an agent would still prefer reporting their true signal, but this won't always be the case. The reported signal only affects payoffs when $\delta_i \geq 1$, so the agent should choose their report $\hat{x}_i$ conditional on this event to maximize the expected score $R(p_{i|\delta_i \geq 1}, q_i(\hat{x}_i))$. Since $p_i \neq p_{i|\delta_i \geq 1}$ in general, this opens up the possibility that the agent would want to misreport their signal if another prediction $p(\hat{x}_i)$ approximates $p_{i|\delta_i \geq 1}$ better than $p_i = p(x_i)$ does under R[3].

For instance, suppose the posterior predictions are $p^a(a) = \mathrm{E}[\bar{x}^a_{-i} \,|\, a] = 0.99$ and $p^a(b) = 0.45$ given two possible signals a and b, with $p^a_{i|\delta_i \geq 1}(a) = \mathrm{E}[\bar{x}^a_{-i} \,|\, a \textit{ and } \delta_i \geq 1] = 0.5$. While an agent with $x_i = a$ thinks B signals are rare, agents with the b signal are relatively better predictors when they are present. Since $R(p_{i|\delta_i \geq 1}(a), p(b)) = R((.5, .5), (.45, .55)) > R((.5, .5), (.99, .01)) = R(p_{i|\delta_i \geq 1}(a), p(a))$ for typical scoring rules, an agent with $x_i = A$ would prefer report $(b, \varnothing)$ over $(a, \varnothing)$.

Despite this possibility, these "failures" of incentive compatibility out of equilibrium are unconcerning. If an agent is sophisticated enough to notice an improvement from misreporting their signal conditional on omitting their prediction, they would be sophisticated enough to best respond by giving a full report. Furthermore, these gains can exist only when $p_{i|\delta_i \geq 1}$ is sufficiently different from p_i for a fixed set of posterior predictions. Since $p_{i|\delta_i \geq 1} \to p_i$ as $\Pr_i[\delta_i \geq 1 \,|\, x_i] \to 1$, this possibility goes away completely for large enough n if agents become increasingly certain that every signal will be given by at least one person along with a prediction.

6. FURTHER CONSIDERATIONS FOR ROBUST MECHANISMS

6.1. Individual Rationality and Budget Balance

In the honest Bayes-Nash equilibrium, each agent receives a score of $R(\bar{x}_{-i}, p_i)$. Depending on the scoring rule R used, these scores could be negative, positive, or sum to any amount. If agents have the option to sit out from the mechanism, then participation shouldn't leave them worse off. Whether agents can't be worse off in every case or on average leads to the following two notions of individual rationality:

Definition 6.1. A truth serum is *ex-post individually rational* (EPIR) if the realized score satisfies $S_i(\boldsymbol{x}, \boldsymbol{p}) \geq 0$ for all i and for all profiles of reports $\boldsymbol{x}$ and $\boldsymbol{p}$.

Definition 6.2. A truth serum is *interim individually rational* (IIR) if the expected score over the types of others conditional on x_i satisfies $\mathrm{E}[S_i((x_i, x_{-i}), (p_i, p_{-i})) \,|\, x_i] \geq 0$ for all i.

Similarly, we can consider two forms of budget-balance depending on whether the total scores for all agents sum to zero for all realizations or on average:

Definition 6.3. A truth serum is *ex-post budget balanced* (EPBB) if the total scores satisfy $\sum_i S_i(\boldsymbol{x}, \boldsymbol{p}) = 0$ for all profiles of reports $\boldsymbol{x}$ and $\boldsymbol{p}$.

Definition 6.4. A truth serum is *ex-ante budget balanced* (EABB) if the total scores satisfy $\mathrm{E}[\sum_i S_i(\boldsymbol{x}, \boldsymbol{p})] = 0$ in expectation over all profiles $(\boldsymbol{x}, \boldsymbol{p})$

Since a primary advantage of the truth serum is that it is detail-free, the notion of ex-ante budget balance isn't applicable. Unfortunately, ex-post budget balance can be

[3] The notion of how close one probability vector is to another under a proper scoring rule R can be formalized as its corresponding *Bregman divergence* $D_R(p, q) = R(p, p) - R(p, q)$. For example, the Bregman divergence of the quadratic scoring rule is squared Euclidean distance, and the Bregman divergence of the log scoring rule is the Kullback-Leibler divergence.

a strong condition. EPBB and EPIR are incompatible for any non-trivial mechanism in this setting. IIR will also tend to be violated under EPBB mechanisms. When agents have common priors, EPBB and IIR will be mutually be satisfied only if the iterim expected scores are exactly zero for each type since there are no gains from interaction. Individual rationality also becomes a more complex goal if agents have some unknown costs of participation, reflecting either effort to acquire a signal [Dasgupta and Ghosh 2013; Witkowski et al. 2013] or concern over privacy [Ghosh and Roth 2013].

6.2. Collusion Resistance and Balanced Truth Serums

As the minimum truth serum is defined, honesty is a Bayes-Nash equilibrium, but it is not the only one. In fact, untruthful equilibria might Pareto-dominate honest revelation. Notice that any strategy profile where the agents draw $\hat{x}_i$ according to a common fixed distribution $\hat{p}$ and report $(\hat{x}_i, \hat{p})$ is a Bayes-Nash equilibrium. Since the expected score is $\mathrm{E}\, S_i(\hat{x}_i, \hat{p}) = R(\hat{p}, \hat{p}) = F(\hat{p})$ for some convex F by the Savage representation, the total scores will be maximized when the agents can coordinate on a degenerate distribution and report some signal t^* with probability one. One benefit of ex-post budget balance is that it guarantees agents cannot coordinate to increase their total score, and hence provides a means of reducing the tempation of any untruthful equilibria.

Define the balanced minimum truth serum as follows:

Definition 6.5 (*Balanced minimum truth serum*). Given a proper scoring rule R and a unanimous aggregator g, the balanced minimum truth serum scores (with or without optional predictions) are

$$S_i^b(\boldsymbol{x}, \boldsymbol{p}) = \frac{1}{n-1} \sum_{j \neq i} \left(S_i(x_{\text{-}j}, p_{\text{-}j}) - S_j(x_{\text{-}i}, p_{\text{-}i}) \right)$$

Since each term $S_i(x_{\text{-}j}, p_{\text{-}j})$ is individually incentive compatible and the terms $S_j(x_{\text{-}i}, p_{\text{-}i})$ don't depend on i's report, these scores are still Bayesian incentive compatible with the caveat that we now require $\Pr_i[n \geq |T| + 2] > 0$ for strict incentive compatibility. Ex-post budget balance follows from each term $S_i(x_{\text{-}j}, p_{\text{-}j})$ appearing in the total scores exactly once with a positive sign—in the score of agent i—and exactly once with a negative sign—in the score of agent j. The uninformative coordination equilibria still exist, but no longer Pareto-dominate the honest equilibrium since the total scores are fixed at zero.

As noted earlier, these mechanisms will not be interim individually rational in general. For example, suppose $T = \{a, b\}$ and $n = 4$. Agents believe signals are conditionally independent based on two states, A and B, with $\Pr[A] = 0.1$, $\Pr[B] = 0.9$, $\Pr[x_j = a \,|\, A] = 0.9$, and $\Pr[x_j = b \,|\, B] = 0.1$. Under the minimum truth serum with the log scoring rule, an agent with $x_i = a$ expects $S_i(x_{\text{-}j}, p_{\text{-}j})$ to be -0.693 and $S_j(x_{\text{-}i}, p_{\text{-}i})$ to be -0.593, leading to the expected balanced score to be approximately -0.1. Thus, agents with $x_i = a$ would prefer a score of zero from not participating if possible.

If the scoring rule R is bounded by the interval $[M_1, M_2]$, then the balanced truth serum scores are always greater than $M_1 - M_2$. Subtracting this negative constant from each agent's balanced score guarantees ex-post individual rationality while making total scores always sum to $n(M_2 - M_1)$.

6.3. Risk Aversion and Probabilistic Rewards

Another point of concern for robustness is the assumption of risk-neutral agents. If agents are risk-averse as we might typically expect, Bayesian incentive compatiblity or interim individual rationality could be violated. When the Bernoulli utility function $u_i(\cdot)$ of an agent is known, assigned scores can be transformed to $u_i^{-1}(S_i(\boldsymbol{x}, \boldsymbol{p}))$ to counteract the agent's risk preference. When the agent's risk preference is unknown, another trick is available: payment in lottery shares. An expected utility maximizer can be risk-averse across varying rewards, but will always have risk-neutral preferences over a varying probability of a fixed reward. Reinterpreting scores as the probability of winning a prize means risk neutrality holds without loss of generality. Now, ex-post individual rationality has a dual role of ensuring scores are proper probabilities. Hossain and Okui [2013] and Schlag and van der Weele [2009] explore this trick theoretically and experimentally in the case of eliciting judgements from a single agent.

There are two ways of incorporating probablistic rewards into a truth serum. First, the mechanism could give each agent the opportunity to win their own prize. Using a scoring rule R bounded in the interval $[0, M]$, the mechanism assigns scores $S_i(\boldsymbol{x}, \boldsymbol{p})$ and draws thresholds $K_i \sim \mathrm{Unif}[0, M]$. Agent i wins their prize if and only if $K_i < S_i(\boldsymbol{x}, \boldsymbol{p})$. Alternatively, the mechanism could award a single prize, splitting shares in the prize lottery according to scores. This entails using a balanced truth serum adjusted so that total scores sum to one and scores are always non-negative:

Definition 6.6 (*Lottery minimum truth serum*). Given a proper scoring rule R bounded between $[0, 1]$ and a unanimous aggregator g, the lottery minimum truth serum scores (with or without optional predictions) are

$$S_i^{\ell}(\boldsymbol{x}, \boldsymbol{p}) = \frac{1}{n} + \frac{1}{n(n-1)} \sum_{j \neq i} \left(S_i(x_{\text{-}j}, p_{\text{-}j}) - S_j(x_{\text{-}i}, p_{\text{-}i}) \right),$$

denoting the probability that agent i wins the prize.

7. CONCLUSION

This paper gives a new class of detail-free mechanisms for eliciting correlated signals. The only restrictive assumption for incentive compatibility is that all agents with the same signal have the same posterior expectations. The obvious next question is whether the common predictions assumption can be weakened, possibly at the cost of different conditions on belief structures. The idea underlying this paper is that under common predictions, another agent with the same signal can act as a perfect proxy for an agent's belief. This suggests incentive compatibility should be feasible when agents think others with the same signal are better proxies on average, if not exactly.

In this paper, I've assumed agents have no direct preferences over how their reports are used, which is plausible for many surveys and ratings. In more general settings, the incentives for honesty provided by a truth serum could be used to counteract other incentives for dishonesty. One direction to explore is when it's possible to layer truth serum transfers on top of another mechanism to provide incentive compatibility. This broad idea has an established history in mechanism design. For instance, Crémer and McLean [1988] employ a similar transfer scheme to prove their full surplus extraction result for correlated types. Though presented more as a paradox than a practical result, their mechanism assumes precise knowledge of the agents' common prior. The question of when detail-free transfers like the minimum truth serum could fill an analogous role in general implementation problems remains open.

REFERENCES

BRIER, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78,* 1, 1–3.

CARVALHO, A. AND LARSON, K. 2012. Sharing rewards among strangers based on peer evaluations. *Decision Analysis 9,* 3, 253–273.

CRÉMER, J. AND MCLEAN, R. 1988. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society 56,* 6, 1247–1257.

DASGUPTA, A. AND GHOSH, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proc. of 22nd Int. World Wide Web Conf. (WWW 2013)*. 319–329.

GHOSH, A. AND ROTH, A. 2013. Selling privacy at auction. *Games and Economic Behavior*.

GNEITING, T. AND RAFTERY, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102,* 477, 359–378.

GOOD, I. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological) 14,* 1, 107–114.

HOSSAIN, T. AND OKUI, R. 2013. The binarized scoring rule. *The Review of Economic Studies 80,* 3, 984–1001.

JURCA, R. AND FALTINGS, B. 2007. Collusion-resistant, incentive-compatible feedback payments. In *Proc. of 8th ACM Conf. on Electronic Commerce (EC 2007)*. 200–209.

JURCA, R. AND FALTINGS, B. 2011. Incentives for answering hypothetical questions. In *Proc. of 1st Workshop on Social Computing and User Generated Content (EC 2011 Workshop)*.

MILLER, N., RESNICK, P., AND ZECKHAUSER, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science 51,* 9, 1359–1373.

PRELEC, D. 2004. A Bayesian truth serum for subjective data. *Science 306,* 5695, 462–466.

RADANOVIC, G. AND FALTINGS, B. 2013. A robust Bayesian truth serum for non-binary signals. In *Proc. of the 27th AAAI Conf. on Artificial Intelligence (AAAI 2013).* 833–839.

SAVAGE, L. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66,* 336, 738–801.

SCHLAG, K. AND VAN DER WEELE, J. 2009. Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. *Theoretical Economics Letters 2013,* February, 38–42.

WITKOWSKI, J., BACHRACH, Y., KEY, P., AND PARKES, D. 2013. Dwelling on the negative: Incentivizing effort in peer prediction. In *Proc. of 1st AAAI Conf. on Human Computation and Crowdsourcing*.

WITKOWSKI, J. AND PARKES, D. 2012a. A robust Bayesian truth serum for small populations. In *Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI 2012)*.

WITKOWSKI, J. AND PARKES, D. 2012b. Peer prediction without a common prior. In *Proc. of the 13th ACM Conference on Electronic Commerce (EC 2012)*.

ZHANG, P. AND CHEN, Y. 2014. Elicitability and knowledge-free elicitation with peer prediction. In *Proc. of the 2014 Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2014)*. 245–252.