Beyond the Bayesian Truth Serum: The Knowledge Free Peer Prediction Mechanism

a thesis presented
by
Peter Zhang
to
The Computer Science and Mathematics Departments

in partial fulfillment of the honors requirements
for the degree of
Bachelor of Arts
in the subject of
Computer Science and Mathematics

Harvard University
Cambridge, Massachusetts
March 2013

Thesis advisor: Yiling Chen Peter Zhang

Beyond the Bayesian Truth Serum: The Knowledge Free Peer Prediction Mechanism

Abstract

The elicitation of private information from individuals is crucially important to many tasks, ranging from scientific research to corporate decision-making. Eliciting private information is particularly challenging when objective truth is inaccessible - when there is no “anwer key” available. To address this challenge, we present the Knowledge Free Peer Prediction mechanism (KFPP). KFPP induces truthful reporting for any number of agents $n \geq 3$, doesn’t require the mechanism to know the common prior, and can handle non-binary information elicitation; it thus improves on previous information elicitation mechanisms designed for this setting, like Peer Prediction, the Bayesian Truth Serum, and the Robust Bayesian Truth Serum. Furthermore, we demonstrate that KFPP can handle several complications, including risk-adverse participants, continuous signals, and participants who experience varying costs when acquiring and reporting their information.

Contents

List of Illustrations **vi**

Figures vi

Tables vi

1 Introduction **1**

1.1 The Setting 1

1.2 Background 2

1.3 Contributions 4

1.4 Outline 5

2 Related Work **7**

2.1 Mechanism Design 7

2.2 Strictly Proper Scoring Rules 11

3 Preliminaries **13**

3.1 The Model 13

3.2 Basic Definitions and Lemmas 15

3.3 Theoretical Motivation 18

4 The Knowledge Free Peer Prediction Mechanism **21**

4.1 Sequential Variant 21

4.2 Simultaneous Variant 26

4.3 Tradeoffs 29

5 Extensions 31
5.1 Risk Adversion 31
5.2 Continuous Signals 32
5.3 Effort Elicitation 34

6 Conclusion 41
6.1 Future Work 41

A Appendix 44
A.1 Appendix 1 44

List of Illustrations

Figures

4.1.1 Sequential game related to the sequential KFPP mechanism. 24

Tables

2.1.1 Key features of KFPP and previous mechanisms. KFPP has all the desirable features of PP without the assumption that the mechanism designer knows the common prior. 10

Acknowledgments

I am deeply grateful to those who made this thesis possible.

First, I'd like to thank my thesis advisor, Yiling Chen, for her insightful guidance and encouragement throughout the process of writing this thesis, as well as her excellent instruction in CS 286r which inspired my interest in this topic.

Second, I'd like to thank my academic advisor and thesis reader, David Parkes, for helpful discussions about the related works and conclusion of this thesis, as well as his help with LaTeXformatting.

I'd also like to thank Michael Mitzenmacher for agreeing to be my thesis reader on such short notice; I look forward to your comments.

Thanks to Carl Jackson, for his keen eye and helpful comments, and to the rest of my blockmates for their emotional support.

Finally, I'd like to thank my parents for their ever-present love and support.

Chapter 1: Introduction

1.1 The Setting

The elicitation of information from individuals is essential to human knowledge-gathering, decision-making, and research. Humans rely on inputs from others for a surprisingly wide range of tasks. When forecasting the future, whether economic, meterological, or otherwise, we often request the analysis of multiple subject-matter experts before forming an outlook. Many academic disciplines, like psychology and sociology, depend heavily on surveys of human perception or emotion to conduct research and further knowledge in the field. Even the weighty task of selecting the leaders of a country requires the solicitation of human opinions, in the form of votes, at least in democratic countries.

The successful completion of all these tasks requires that the human respondents being queried are truthful about their private information; dishonest votes, survey answers, or expert opinions can lead to unrepresentative elections, faulty research results, and inaccurate forecasts. Unfortunately, humans are naturally self-interested, and often have incentives to report dishonestly, or not report at all. When participants are interested in the ultimate outcome of the information elicitation process, such as an election or an auction, it is generally relatively easy to encourage them to report their preferences truthfully - there are well-known voting systems and auction designs that are incentive compatible. However, in many situations, information gathering is beneficial, but the ultimate outcome of the process does not directly impact the subjects from whom the information is being elicited. When this is the case, how can we ensure that participants are still reporting honestly?

If the information being elicited is objective and easy to verify, then the application of *strictly proper scoring rules* can be used to align the incentives of information-providing agents with truthful reporting. The core idea behind scoring rules is to grade the submitted information against either an objective answer key, or the realization of some related public event in the future, and reward the information-providers in such a way that they have an economic incentive to report their information truthfully. For example, when gathering forecasts about tomorrow's weather, the information-gatherer could wait a day, and reward experts based on how accurately their predictions describe the actual weather at that time. As long as there is some way for information-providers to be graded against some objective and publically available standard, it is a relatively simple task to create truth-promoting incentives.

However, when an "answer key" is unavailable, either because the information is inherently subjective - like personal opinions or emotions - or because the information is objective but practically difficult to observe for an outsider - like the number of hours a telecommuting employee has actually worked - such techniques are no longer directly applicable. It is this setting that we are primarily interested in: eliciting information when objective truth is not accessible.

1.2 Background

Over the last decade, there have been several significant contributions to this area of research: the Peer Prediction method (PP), the Bayesian Truth Serum (BTS), and the Robust Bayesian Truth Serum (RBTS). All three of these mechanisms model the task of information elicitation in the same basic manner: the participants of the mechanism each receive a signal, or private information, from the world, and all participants share the same common prior belief about the signal distribution over the participants. We will discuss each of these mechanisms in turn.

The Peer Prediction method, proposed by Miller *et al.* (2005), was the first mechanism designed specifically to address incentive alignment for honest reporting of information when "independent, objective outcomes are not available." The central idea behind this method is the application of a strictly proper scoring rule on one participant's signal report, based on how well it predicts another participant's report. Simply comparing two participants' reports and rewarding agreement is problematic for participants who have signals that they believe to be rare, as they would increase their expected payout by falsely reporting a more common signal. Instead, Peer Prediction formulates a predictive distribution over the report of a second participant based on the signal report of the first participant, and rewards the first participant based on how accurately this prediction matches the actual report by the second participant, as measured by a strictly proper scoring rule. In this way, Peer Prediction deftly circumvents the lack of an objective truth; it is a strict Bayes-Nash equilibrium for all participants to report their signal truthfully. However, this method requires that the mechanism designer (the person who is soliciting signal reports) knows the common prior of the participants, so that it can accurately calculate the posterior distribution to score after receiving a participant's signal. Unfortunately, this assumption is quite strong in many contexts. For example, when a mechanism designer asks a question for the first time, he often has no knowledge about the probabilistic relationship between the signals; more generally, the less familiar the mechanism designer is with the problem space, the more unlikely it is that he is familiar with the beliefs of the participants of the mechanism.

To address this issue, Prelec (2004) proposed the Bayesian Truth Serum. Unlike Peer Prediction, the BTS asks participants to report not only their own signal, but also predict the distribution of signals for the whole population (all the participants). To ensure truthful reporting of this distribution, the mechanism designer gives everyone a *prediction score* dependent on how well their predicted distribution matches the realized signal distribution, as elicited by the mechanism. Then, the mechanism designer rewards participants for their signal reports with a *information score*, based on how *surprisingly common* their reported signal is in the realized signal distribution, as compared to the concensus predicted signal distribution. The combination of these two scores is translated into a monetary reward for the participant. Intuitively, a participant believes that her own signal will be surprisingly common, since other participants who did not receive the same signal will predict a mistakenly low frequency for that signal, and is thus incentivized to report her signal truthfully. In this way, the BTS aligns incentives for truthful reporting of signals and signal frequencies - truthful reporting is again a strict Bayes-

Nash equilibrium - without relying on knowledge of the common prior. However, the BTS suffers from two major flaws. First, truthful reporting is only incentive compatible given a large number of participants, and this number is dependent on the common prior. Thus, in practice, a mechanism designer who does not know the common prior cannot make the BTS truly incentive compatible; furthermore, even if the mechanism designer does have some knowledge of the common prior, he may find it difficult to recruit a sufficient number of participants to meet the requirements for incentive compatibility. Second, the BTS is not ex-post individually rational, meaning that it does not guarantee non-negative payouts to all participants. It may be infeasible for a mechanism designer to demand payments from a participant in a real-world application of this mechanism.

Witkowski and Parkes (2012b) improved upon the Bayesian Truth Serum by proposing the Robust Bayesian Truth Serum. Like the BTS, the RBTS requires participants to make both an information and a prediction report. However, rather than reward information reports that are *surprisingly common*, the RBTS rewards participants based on how well their information report can be used to update the prediction report of another participant using the *shadowing* technique. On the one hand, the use of this technique guarantees that honest reporting is a strict Bayes-Nash equilibrium for any number of participants ≥ 3, and guarantees all participants receive non-negative payouts. On the other hand, as Witkowski and Parkes (2012b) notes, the RBTS can only elicit binary information; as we'll demonstrate in Chapter 3, the *shadowing* technique does not extend nicely to the elicitation of non-binary information.

1.3 Contributions

In this thesis, we present the Knowledge Free Peer Prediction (KFPP) mechanism. The KFPP takes the same reports as the RBTS (and the BTS) - an information report and a prediction report - and retains the desirable properties of RBTS. In particular, it is ex-post individually rational, and there is a strict equilibrium where all agents report their signal truthfully, for any number of agents ≥ 3. The primary advantage of KFPP as opposed to RBTS is that it is capable of handling non-binary signals.

The main innovation of KFPP is a technique that allows the mechanism designer to properly formulate a posterior distribution based on an agent's reported signal, as is necessary in the Peer Prediction method, without actually knowing the common prior. Specifically, it outsources the task of bayesian updating to the participants themselves - this is possible we assume the common prior exists, and thus all agents know the common prior - and incentivizes them to do so truthfully with a prediction score. In this way, KFPP solves the fundamental weakness of the Peer Prediction method without limiting the applicability of the mechanism, unlike BTS and RBTS.

In addition, we demonstrate that KFPP, like Peer Prediction, can easily handle continuous signals and risk-adverse agents, using analogous methods to those proposed by Miller *et al.* (2005). We also address the possibility of agents incurring unknown costs when acquiring and reporting a signal, which may threaten the mechanism designer's ability to elicit effort. We prove that it is impossible to construct a mechanism that is both ex-post individually rational and has an equilibrium where all participants report their signal truthfully. Finally, we proceed to construct a mechanism that is interim individually rational and has an equilibrium where any number of participants (other than the total number of participants) report their signal truthfully, by combining KFPP with a uniform auction.

1.4 Outline

The remainder of this thesis are organized as follows:

- Chapter 2 covers other related work in addition to those already mentioned in the introduction.
- In Chapter 3, we provide some background knowledge necessary to understand and motivate KFPP.
- In Chapter 4, we formally describe two variants of KFPP - a sequential and a non-sequential version - and prove that the former has a strict Perfect Bayesian equilibrium and the latter has a strict Bayes-Nash equilibrium where all participants report their signals truthfully, and that both variants are ex-post individually rational.

- Chapter 5 discusses several extensions to the KFPP mechanism, including the handling of continuous signal spaces, risk-adverse agents, and agents that experience unknown costs when acquiring and reporting their signal.
- We conclude in Chapter 6 with discussion of potential avenues for future work.

Chapter 2: Related Work

2.1 Mechanism Design

The theoretical foundation for the work in this area is a subfield of game theory called *mechanism design*. Sometimes called reverse game theory, *mechanism design* is primarily interested in the design of games, or mechanisms, with certain desirable equilibria, rather than the equilibrium analysis of a specific game. Hurwicz, Maskin, and Myerson were awarded the Nobel Prize in Economics in 2007 for founding the field.

First, we give formal descriptions of the three predecessor mechanisms discussed in Chapter 1. Let n denote the number of agents. Agents have information in the form of a private signal received from the world; denote the space of possible signals be $\mathcal{S} = \{s_1, ...s_o\}$. Furthermore, let R_p denote any strictly proper scoring rule, and R_q denote the quadratic scoring rule; see Chapter 3 for definitions of these terms.

Peer Prediction

Every agent i is asked to report their signal; let $x_i \in \mathcal{S}$ be the report of agent i. Let p denote the common prior belief about the distribution of the signals, and let p_s denote the posterior belief about the distribution of the signals with knowledge of signal $s \in \mathcal{S}$. For every agent i, select a reference agent $j \neq i$. Agent i receives a payout of

$$R_p(p_{x_i}, x_j)$$

Under the Peer Prediction method, it is a Bayes-Nash equilibrium for every agent to report their signal truthfully, as shown in Miller *et al.* (2005). Note that this is the case only if one agent's signal is *stocastically relevant* to another agent's signal - the distribution of a second agent's signal conditional on the first agent's signal is different for different realizations of the first agent's signal - which somewhat restricts the class of acceptable common priors.

Bayesian Truth Serum

Every agent i is asked for two reports:

- Information report: Let $x_i \in \mathcal{S}$ be agent i's reported signal.
- Prediction report: Let $\vec{y_i} = (y_i^1, ...y_i^o)$ be agent i's report about the frequency of the signals, with y_i^j being agent i's prediction about the frequency of signal s_j.

For every signal s_j, define

$$f_j(x_i) = \begin{cases} 1 & \text{if } x_i = s_j \\ 0 & \text{otherwise} \end{cases}$$

Then, for every signal s_j, the mechanism designer computes

$$\overline{x}_{s_j} = \frac{1}{n} \sum_{k=1}^{n} f_j(x_k)$$

$$\log \overline{y}_{s_j} = \frac{1}{n} \sum_{k=1}^{n} \log y_k^j$$

Finally, the payout for agent i is

$$\underbrace{\log \frac{\overline{x}_{x_i}}{\overline{y}_{x_i}}}_{\text{information score}} + \underbrace{\alpha \sum_{k=1}^{o} \overline{x}_{s_k} \log \frac{y_i^k}{\overline{x}_{s_k}}}_{\text{prediction score}}$$

for some $\alpha > 0$. Prelec (2004) demonstrates that the Bayesian Truth Serum has a strict Bayes-Nash equilibrium where all agents report truthfully, as long as two conditions hold. First, the number of agents must be sufficiently large; the exact number depends on the common prior. Second, agents must treat signals as *impersonally informative* about the population signal distribution; every agent must believe that all other agents who received the same signal have the same belief about the population signal dstribution. A formal description of the modeling assumptions made by is the BTS is given in Chapter 3; KFPP uses the same model.

Robust Bayesian Truth Serum

The Robust Bayesian Truth Serum can only handle binary single spaces. WLOG, we will call the two signals 1 - the high signal - and 0 - the low signal. 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 the high signal.

Next, for each agent i, select a reference agent $j = i+1 \mod n$ and a peer agent $k = i+2 \mod n$. The mechanism designer then calculates the following for every agent:

$$\delta = \min(y_j, 1 - y_j)$$

$$y_i' = \begin{cases} y_j + \delta & \text{if } x_i = 1 \\ y_j - \delta & \text{if } x_i = 0 \end{cases}$$

Finally, agent i is given a payout of

$$\underbrace{R_q(y_i', x_k)}_{\text{information score}} + \underbrace{R_q(y_i, x_k)}_{\text{prediction score}}$$

The Robust Bayesian Truth Serum has a Bayes-Nash equilibrium where all agents report truthfully, and guarantees non-negative payouts for all agents, given any number of agents

Mechanism	Signal space			Common prior ...		Incentive Compatible for ...	Ex-Post Individually Rational
	bin.	all discrete	cont.	is known to mechanism	exists		
PP	✓	✓	✓	✓	✓	≥ 2	✓
BTS	✓	✓			✓	?	
RBTS	✓				✓	≥ 3	✓
BPP & SPP	✓					≥ 3	✓
KFPP	✓	✓	✓		✓	≥ 3	✓

Table 2.1.1: Key features of KFPP and previous mechanisms. KFPP has all the desirable features of PP without the assumption that the mechanism designer knows the common prior.

≥ 3, as shown by Witkowski and Parkes (2012b). This holds under the same assumptions as BTS, as described in Chapter 3.

In addition to these mechanisms there is another related work that directly addresses the problem of eliciting information when objective truth is not accessible. Witkowski and Parkes (2012a) introduces two related mechanisms, Basic Private-Prior Peer Prediction (BPP) and Shadow Private-Prior Peer Prediction (SPP). In addition to the shadowing technique used by Witkowski and Parkes (2012b), these mechanisms depend on the concept of *temporal separation* - the assumption that the mechanism designer has access to the participants both before and after they receive their signal. BPP and SPP go further than BTS and RBTS in improving the Peer Prediction method by not only removing the assumption that the mechanism designer knows the common prior, but also removing the assumption that the common prior exists at all; BPP and SPP can handle situations where the participants have different prior beliefs about the signal distribution. A comparison between the key features of these mechanisms and KFPP can be found in table 2.1.1.

In our analysis of effort-elicitation from agents who experience cost, we use two additional results from mechanism design.

The first is the *Revelation Principle.* Introduced by Gibbard (1973) for dominant strategy equilibria, and later extended to Bayesian Equilibria by Myerson (1979), the Revelation Principle states that for any mechanism with some equilibrium, there exists a direct-revelation mechanism with a pay-off equivalent equilibrium where all agents report their type truthfully. We will use the *Revelation Principle* to prove that it is impossible to construct a mechanism that is both interim individually rational and has an equilibrium where all participants report their signal truthfully.

The second is the *uniform price auction*, one possible extension of a second-price auction. While it does not guarantee truthful reporting of type in general, a *uniform price auction* is incentive compatible when all bidders have demand for exactly one unit of the good being auctioned, as is the case when we apply it. We use this type of auction to construct a mechanism that is interim individually rational and has an equilibrium where any number of participants (other than the total number of participants) report their signal truthfully. For a technical analysis of this auction, see Krishna (2002, p. 190–196).

2.2 Strictly Proper Scoring Rules

The concept of a strictly proper scoring rule was first introduced by Brier. (1950) for the purpose of verifying meterological forecasts. While Brier doesn't use the term scoring rule, he provides a formula for scoring predictions of a future event (in the form of a discrete probability distribution over the possible realizations of the event) which is uniquely maximized when the predictor predicts her true belief. Generally, strictly proper scoring rules are restricted to scoring discrete probability distributions; however, some work has been done to extend them to scoring continuous distributions. Many popular discrete strictly proper scoring rules, like the quadratic, logarithmic, and spherical scoring rules, have simple continuous analogs, but their continuous analogs are not defined for all continuous probability distributions. Matheson and Winkler (1976) describes a technique for deriving continuous scoring rules from binary scoring rules that do not suffer from this flaw.

While strictly proper scoring rules in and of themselves are only applicable to information elicitation when objective truth is accessible, they are a significant component in many mechanisms designed to elicit information when objective truth is not accessible - Peer Prediction, the Robust Bayesian Truth Serum, Basic Private-Prior Peer Prediction and Shadow Private-Prior Peer Prediction all use scoring rules. We use discrete strictly proper scoring rules to construct the Knowledge Free Peer Prediction mechanism, and continuous strictly proper scoring rules to extend KFPP to handle continuous signal spaces.

For a more recent treatment of strictly proper scoring rules, along with a formal characterization of all strictly proper scoring rules, see Gneiting and Raftery (2007). We also discuss strictly proper scoring rules in more depth in Chapter 3.

Chapter 3: Preliminaries

Before we introduce the KFPP mechanism, we provide some preliminary background. In section 3.1 we specify the assumptions we make when modeling the problem of eliciting information from others. Next, we cover some definitions and lemmas necessary for the construction and analysis of KFPP in section 3.2. Finally, in section 3.3, we give a theoretical motivation for KFPP by demonstrating that the *shadowing* technique used in RBTS does not extend nicely to situations with non-binary signals. Many of definitions, lemmas, and modeling assumptions are either drawn directly from or inspired by Witkowski and Parkes (2012b).

3.1 The Model

We model the problem of eliciting information from others as follows. There are $n \geq 3$ rational, risk-neutral agents who seek to maximize their expected payout. All agents share the same probabilistic belief system about the structure of the world, which consists primarily of states and signals. There are m possible world states; the true world state will be represented by the random variable T, which resolves to a value in $\{1, ..., m\}$, but is never observed by any agent. Each agent i receives a signal represented by the random variable $S_i \in \{s_1, ...s_o\}$, which may represent their experience, opinion, or other private information; we denote a generic signal by S. Our ultimate goal is to elicit the true signal that each agent received. All agents share a common prior, which consists of a shared prior distribution over the world state $P(T = t)$ and a shared belief about $P(S = s_i|T = t)$, the probability of receiving a particular signal conditional on the world state. Upon receiving a signal s_{z_i}, agent i can update her posterior belief $P(S_j = s_k|S_i =$

s_{z_i}) that agent j receives some signal s_k as follows:

$$P(S_j = s_k | S_i = s_{z_i}) = \sum_{t=1}^{m} P(S_j = s_k | T = t) P(T = t | S_i = s_{z_i}),$$

where $P(T = t | S_i = s_{z_i})$ can be computed using Bayes' rule. Note that because we assume that the probability of receiving a particular signal is only dependent on the world state (it is independent of the identity of the agent receiving the signal) and that all agents share the same common prior, we can denote the generic posterior belief of an agent with knowledge of a signal s_b on the probability that another agent receives the signal s_a as follows:

$$p^{s_a}_{\{s_b\}} = P(S_j = s_a | S_i = s_b) \text{ for any } i \neq j.$$

Similarly, we will extend this notation to "second order" posteriors as well. Let

$$p^{s_a}_{\{s_b, s_c\}} = P(S_k = s_a | S_i = s_b, S_j = s_c) \text{ for any } i \neq j \neq k$$

denote the generic posterior belief that any agent with knowledge of two signals s_b and s_c has about the probability of another arbitrary agent receiving signal s_a. More generally, we denote the generic posterior belief of any agent with knowledge of a signal s_b about the signal distribution of another arbitrary agent by $\vec{p}_{\{s_b\}}$, and the generic posterior belief of any agent with knowledge of two signals s_b and s_c about the signal distribution of another arbitrary agent by $\vec{p}_{\{s_b, s_c\}}$.

We will limit our attention to a certain class of common priors, which we call *admissible*.

Definition 3.1.1. *The common prior is admissible if it satisfies the following properties:*

1. $m \geq 2$ *(there are two or more possible world states).*

2. $P(T = t) > 0$ *for all* t *(every state has positive probability).*

3. *There exists a* s_i *such that* $P(S = s_i | T = t) \neq P(S = s_i | T = t')$ *for every* $t \neq t'$ *(states are distinct).*

4. *$1 > P(S = s_i|T = t) > 0$ for all i and t (the signal beliefs, conditional on world state, are fully mixed).*

5. *$\vec{p}_{s_a,s_b} \neq \vec{p}_{s_a,s_c}$ for any $b \neq c$ and any a (stocastic relevance).*

Note that requirements (2) and (3) do not functionally limit the space of common priors; world states with zero probability can be dropped and world states that are identical can be merged without any change to the agents' prior or posterior beliefs about the signal distribution. Requirement (1) is necessary for a signal to be informative; if all agents believe there is only one world state, then an agent's prior and posterior beliefs about the distribution of signals will be identical. Requirement (5) is analogous to the stochastic relevance assumption in Peer Prediction; we differ from the model assumed by BTS and RBTS with this requirement. Requirements (4) and (5) are the strongest constraints, but are necessary to guarantee that the truthful equilibrium is strict, as shown in Chapter 4; without them, truthful reporting is still an equilibrium, just not a strict one.

3.2 Basic Definitions and Lemmas

In addition to the notation introduced in the previous section, there are a few more definitions and concepts that are necessary for understanding and analysing Knowledge Free Peer Prediction. First, we cover scoring rules in section 3.2.1, and then introduce some concepts and terminology from mechanism design in section 3.2.2.

3.2.1 Scoring Rules

Definition 3.2.1. *Given an outcome space $\mathcal{O}$ and $\mathcal{P}$, the class of valid probability distributions over the outcome space $\mathcal{O}$, a **scoring rule** is a function*

$$S : \mathcal{P} \times \mathcal{O} \to \mathbb{R}$$

Informally, a **scoring rule** is a function that grades a forecast of an event against the actual outcome of the event. We are particularly interested in a specific class of scoring rules.

Definition 3.2.2. *A* ***strictly proper scoring rule*** *S is a scoring rule that satisfies the following property. For any $P, P' \in \mathcal{P}$ such that $P \neq P'$,*

$$\mathbb{E}_P[S(P, O)] > \mathbb{E}_P[S(P', O)].$$

where O is a random variable representing a realized outcome from $\mathcal{O}$.

In particular, the best response for any rational agent with belief $P \in \mathcal{P}$ about $\mathcal{O}$ who wishes to maximize her expected score under a strictly proper scoring rule is to report her true belief P. We will denote a generic strictly proper scoring rule by R_p from this point forward.

One strictly proper scoring rule for discrete outcome spaces is the quadratic scoring rule.

Definition 3.2.3. *Consider an outcome space $\mathcal{O} = \{o_1, ..., o_m\}$ consisting of m mutually exclusive events, and a probability distribution $\vec{p} \in \mathcal{P}$ on $\mathcal{O}$. Denote the actual outcome to be $o \in \mathcal{O}$. Then, the* ***quadratic scoring rule*** *is*

$$R_q(\vec{p}, o) = 2p_o - \sum_{i=1}^{m} p_{o_i}^2$$

where p_o represents the probability assigned to outcome o under distribution $\vec{p}$.

For a proof of the strictly proper nature of the quadratic scoring rule, see Selten (1998). The quadratic proper scoring rule has two nice properties. First, since the strictness of a scoring rule is preserved under affine transformation, the quadratic scoring rule can easily be transformed to only produce non-negative scores, by adding m to every score, where m is the number of outcomes in the outcome space. In addition, we observe the following Lemma, also proved by Selten (1998).

Lemma 1. *(Selten, 1998) Let $\vec{p}$ be your true belief about the probability distribution over an event $\mathcal{O}$ with m distinct outcomes. The expected score loss of reporting $\vec{r}$ instead* *of your true belief under the quadratic scoring rule is proportional to the square of the Euclidean distance between the two:*

$$||\vec{p} - \vec{r}||^2 = \sum_{i=1}^{m}(p_i - r_i)^2$$

In other words, given a set of reports $\{\vec{r_i}\}$ to choose to submit to the quadratic scoring rule, an agent maximizes her expected utility (minimizes her expected loss) by selecting the report that is closest to her true beliefs, in terms of Euclidean distance.

3.2.2 Mechanism Design

In this section, we cover some terminology relating to mechanism design that we adopt for the remainder of this thesis.

Definition 3.2.4. *A mechanism is **ex-post individually rational** if it guarantees non-negative payments to all agents.*

An example of an **ex-post individually rational** mechanism is RBTS. Intuitively, the concept of individual rationality is meant to capture a participant's willingness to participate in the mechanism; they only desire to do so if they believe it gives them non-negative payout. However, this type of individual rationality is quite strong. We introduce a slightly weaker form of individual rationality.

Definition 3.2.5. *A mechanism is **interim individually rational** if all agents believe that their expected payout from the mechanism is non-negative after they have received their type (in the equilibrium being implemented by the mechanism).*

The **interim individual rationality** criterion ensures that all agents wish to participate in the mechanism even after they know their own type, because their payout is non-negative in expectation; however, the mechanism may make negative payments to agents with non-zero probability.

Next, we introduce the concept of incentive compatibility, which indicates that a mechanism enforces truthful reporting.

Definition 3.2.6. *A simultaneous mechanism is* ***strictly Bayes-Nash incentive compatible*** *if it is a strict Bayes-Nash Equilibrium for all agents to report their signal truthfully under the mechanism.*

For example, RBTS is strictly Bayes-Nash incentive compatible for $n \geq 3$, a binary signal space, and all admissible priors. For a sequential mechanism, we define the following related term.

Definition 3.2.7. *A sequential mechanism is* ***strictly Perfect Bayesian incentive compatible*** *if it is a strict Perfect Bayesian Equilibrium for all agents to report their signal truthfully under the mechanism.*

3.3 Theoretical Motivation

To motivation KFPP, we will demonstrate that the technique used by the Robust Bayesian Truth Serum, *shadowing*, does not extend nicely to a non-binary signal space. Consider the most intuitive extension of the Robust Bayesian Truth Serum to a signal space $S = \{s_1, ...s_o\}$ with cardinality $o > 2$. Just as before, every agent i is asked for two reports:

- Information report: Let $x_i \in \{s_1, ...s_o\}$ be agent i's reported signal.
- Prediction report: Let $\vec{y_i} = (y_i^1, ...y_i^o)$ be agent i's report about the frequency of the signals, with y_i^j being agent i's prediction about the frequency of signal s_j.

Next, for each agent i, select a reference agent $j = i+1 \mod n$ and a peer agent $k = i+2 \mod n$. The mechanism designer then calculates the following for every agent:

$$\delta = \min(\min_i(y_j^i), \min_i(1 - y_j^i))$$

$$\vec{\delta_i} = (-\delta/(o-1), ..., -\delta/(o-1), \overbrace{\delta}^{i^{\text{th}} \text{ entry}}, -\delta/(o-1), ..., -\delta/(o-1))$$

$$\vec{y_i}' = \begin{cases} \vec{y_j} + \vec{\delta_1} & \text{if } x_i = s_1 \\ \vec{y_j} + \vec{\delta_2} & \text{if } x_i = s_2 \\ \vdots & \\ \vec{y_j} + \vec{\delta_o} & \text{if } x_i = s_o \end{cases}$$

Finally, agent i is given a payout of

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

Just as in RBTS on a binary signal space, we shadow the reference agent j's prediction report using agent i's information report and score this adjusted prediction against the peer agent's signal. The constant we shadow by, δ_i, is chosen so that it increases the probability assigned to the signal specified by agent i's information report (but does not increase that probability to greater than 1), and decreases the probability assigned to all the other signals uniformly, so that $\vec{y_i}'$ is still a valid probability distribution; this is analogous to the way δ is chosen in the original RBTS.

Theorem 1. *The natural extension of the RBTS to non-binary signal spaces is not strictly Bayes-Nash incentive compatible for all admissible priors and any number of agents $n \geq 3$.*

Proof. We provide a situation for which the natural extension of RBTS is not strictly Bayes-Nash incentive compatible. Consider three agents sharing the following common prior with $m = 2$ states and $o = 3$ signals:

Ω	s_1	s_2	s_3
$P(T = 1) = 0.5$	$P(s_1 \mid T = 1) = 0.1$	$P(S = s_2 \mid T = 1) = 0.1$	$P(S = s_3 \mid T = 1) = 0.8$
$P(T = 2) = 0.5$	$P(s_1 \mid T = 2) = 0.4$	$P(S = s_2 \mid T = 2) = 0.5$	$P(S = s_3 \mid T = 2) = 0.1$

Consider agent $i = 1$; her reference agent is $j = 2$ and peer agent is $k = 3$. Assume that agent i received signal s_1. Then, if agents j and k are reporting truthfully, agent i strictly prefers reporting signal s_2 to reporting s_1 as her information report, regardless of what signal j received. Accordingly, truthful reporting cannot be a Bayes-Nash equilibrium in this situation, since agent i has a profitable deviation. Thus, RBTS is not strictly Bayes-Nash incentive compatible. □

The intuition behind this counter-example is as follows. After receiving signal s_1, agent i's assessment of the probability that agent k's signal is s_1 increases; however her assessment of the probability that agent k's signal is s_2 increases even more. This is due to the fact that receiving signal s_1 increases agent i's belief that the current world state is $T = 2$, but s_2 is even more likely in that world state than s_1 is. Accordingly, agent i will prefer to shadow towards s_2 rather than towards s_1, and will thus benefit from misreporting.

While the natural extension of the RBTS is not strictly Bayes-Nash incentive compatible for all admissible common priors, it turns out that it works for many reasonable common priors (see appendix A.1 for more details). Nevertheless, the failure of RBTS to extend generally to non-binary signal spaces motivates KFPP.

Chapter 4: The Knowledge Free Peer Prediction Mechanism

In this chapter, we give a formal description of two variants of the KFPP mechanisms, a sequential mechanism and a simultaneous mechanism. In section 4.1, we discuss the sequential mechanism, and prove that it strictly Perfect Bayesian incentive compatible and ex-post individually rational. In section 4.2 we discuss the simultaneous mechanism, and prove that it is strictly Bayes-Nash incentive compatible and ex-post individually rational. Finally, in section 4.3 we discuss the advantages and disadvantages of both variants of KFPP.

4.1 Sequential Variant

4.1.1 Mechanism

For every agent i, select two reference agents $h = i - 1 \mod n$, $j = i + 1 \mod n$, and a peer agent $k = i + 2 \mod n$. Now, all n players play the following sequential game:

1. Round 1: Every player simultaneously reports his signal $x_i \in \{s_1, ..., s_o\}$ to the mechanism.

2. Round 2: Every player i receives the report of player h, x_h, from the mechanism, and then reports the frequency of the signals $\vec{y_i} = (y_i^1, ..., y_i^o)$.

At the end of the game, player i receives payoff

$$\underbrace{R_p(\vec{y}_j, x_k)}_{\text{Information Score}} + \underbrace{R_p(\vec{y}_i, x_j)}_{\text{Prediction Score}}$$

where R_p is any strictly proper scoring rule.

4.1.2 Equilibrium Analysis

The central idea behind the sequential variant of KFPP is that the mechanism designer outsources any bayesian updating necessary to perform Peer Prediction, and thus doesn't need to know the common prior. However, simply having all agents report their signal, and their posterior signal distribution based on the signal, is not sufficient. While we can using a strictly proper scoring rule to induce truthful reporting of the signal distribution, assuming all agents truthfully report their signal, agents have no incentive to report their signal truthfully; accordingly, truthful reporting will be a Bayes-Nash equilibrium, but not a strict Bayes-Nash equilibrium. To motivate agents to report their signal truthfully, we pass their signal report to another agent and let that agent perform further updating based on this report; the final second-order posterior distribution is then graded with a scoring rule. In this way, an agent's signal report is actively being used to update another agent's signal distribution report, and thus all agents have an incentive to report their signal truthfully, just as in the Peer Prediction method. We provide a formal proof of the incentive compatibility and ex-post incentive rationality of the mechanism below.

Theorem 2. *The sequential variant of the KFPP mechanism is strictly Perfect Bayesian incentive compatible for all admissible common priors and any $n \geq 3$ agents.*

Proof. We wish to show that it is a strict Perfect Bayesian Equilibrium for all agents to play truthfully - to report their signal truthfully in the first round and to report $\vec{p}_{\{s_{z_i}, x_h\}}$, the generic posterior belief given information about the two signals s_{z_i} (player i's signal) and x_h (player h's reported signal), in the second round. First, consider the following related game:

1. Player 1 reports his signal $x_1 \in \{s_1, ..., s_o\}$ - not necessarily honestly - to player 2.

2. Player 2 reports the frequency of the signals $\vec{y_2} = (y_2^1, ..., y_2^o)$ after receiving player 1's signal.

The payout for both players at the conclusion of this game is $R_p(\vec{y_2}, s)$, where s is the signal of some other player. We will show that it is a strict Perfect Bayesian Equilibrium for both players to play truthfully in this game, i.e. for player 1 to report his signal truthfully and for player 2 to report $\vec{p}_{\{x_1, s_{z_2}\}}$.

To see this, consider the extensive form representation of this game in figure 4.1.1 below. First, Nature assigns a signal in $\{s_1, ..., s_o\}$ to player 1 (it assigns a signal to player 2 as well, but this isn't relevant to our analysis) according to the common prior. Next, player 1 observes his own signal, and reports a signal to player 2. Player 2 observes the signal that player 1 reports, but cannot directly observe the signal player 1 received and thus cannot tell if player 1 played truthfully; accordingly, player 2 has o information sets, each corresponding to the histories of the game where Nature has assigned any signal to player 1, and player 1 has played the signal s_i for some $s_i \in \{s_1, ..., s_o\}$. We will denote the information set where player 1 has played s_i by I_i. Now player 2 reports a probability distribution $\vec{p}$ on signals. Theoretically, player 2 has a continuum of plays (reporting any value in $\vec{p} \in [0, 1]^o$ such that $\sum_{i=1}^{o} p_i = 1$), but to simplify the game tree we only consider two possibilities: reporting $\vec{p}_{\{x_1, s_{z_2}\}}$ (playing T), or reporting any other probability distribution $\vec{p'} \neq \vec{p}_{\{x_1, s_{z_2}\}}$ (playing F). We will show that playing T is better than playing F for any $\vec{p'}$ in our equilibrium.

We claim that the following assessments constitute a strict Perfect Bayesian Equilibrium for this game:

1. Player 1: Play s_i at node (s_i) for any $i \in \{1, ..., o\}$, with the trivial belief on all information sets, as they are all singletons.

2. Player 2: Play T at all nodes, with the belief that the current history is (s_i, s_i) with 100% probability in information set I_i for all i.

First, note that the beliefs of both players are consistent. In particular, player 2's beliefs are derived directly from Bayes rule applied to player 1's strategy profile, and there are

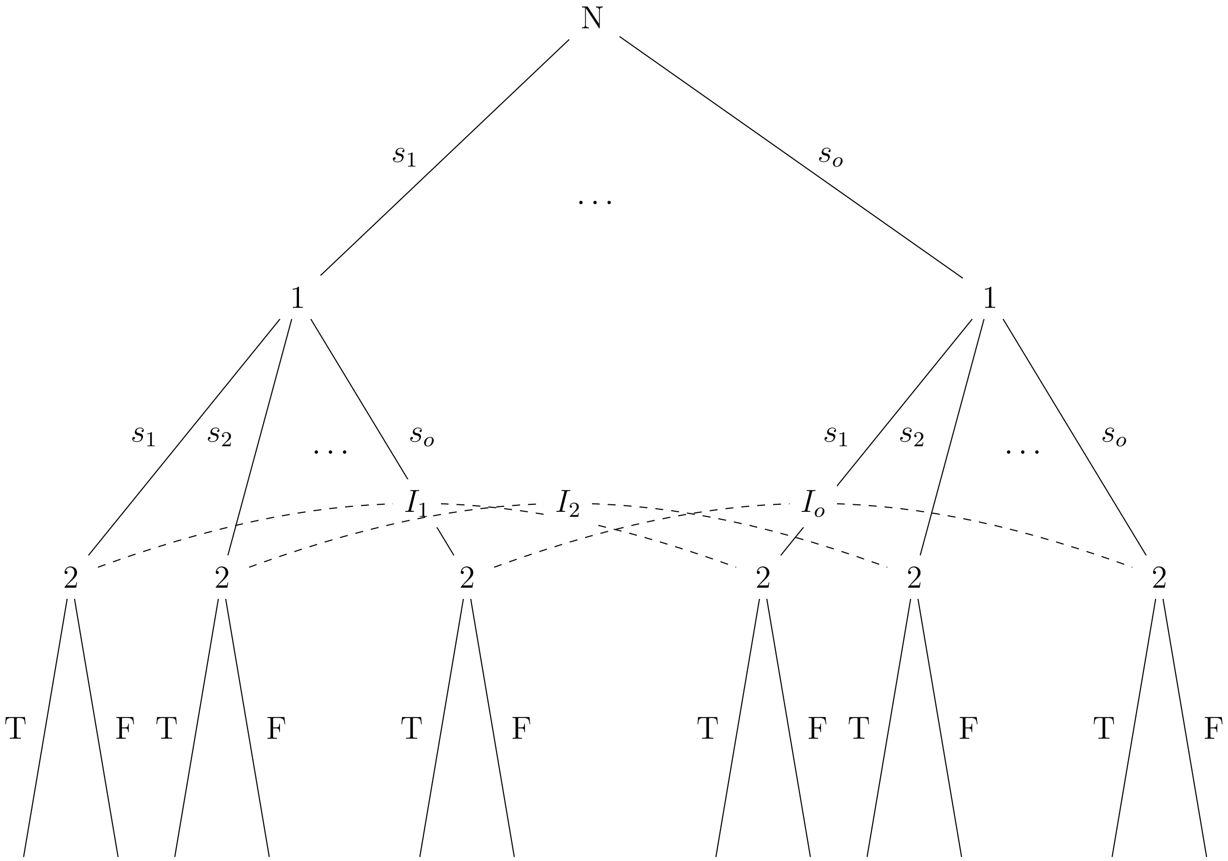

Figure 4.1.1: Sequential game related to the sequential KFPP mechanism.

no off-equilibrium paths, since the common prior is admissible, and thus Nature assigns every signal to player 1 with nonzero probability.

Next, we will show that the assessments are sequentially rational. First, consider any information set of player 1; denote the signal received by player 1 by s_{z_1}. Since R_p is strictly proper, player 1 knows that a report of $\vec{p}_{\{s_{z_1}, s_{z_2}\}}$, conditional on player 2's signal s_{z_2}, will uniquely maximize her expected payout. Accordingly, player 1 has no profitable deviation. Playing any $s' \neq s_{z_1}$ will result in a payout of $R_p(\vec{p}_{\{s', s_{z_2}\}})$, but $\mathbb{E}[R_p(\vec{p}_{\{s', s_{z_2}\}})] < \mathbb{E}[R_p(\vec{p}_{\{s_{z_1}, s_{z_2}\}})]$ since $\vec{p}_{\{s', s_{z_2}\}} \neq \vec{p}_{\{s_{z_1}, s_{z_2}\}}$. Accordingly, no mixed strategy will have a higher payout either, since it will have an expected payout $\leq \mathbb{E}[R_p(\vec{p}_{\{s_{z_1}, s_{z_2}\}})]$, with strict inequality if the mixed strategy involves playing any $s' \neq s_{z_1}$ with positive probability. Next, consider any information set I_i of player 2. At I_i, player 2 believes with certainty that he is at (s_i, s_i). Accordingly, conditional on this additional signal information, player 2 knows that reporting $\vec{p}_{\{s_i, s_{z_2}\}}$ will uniquely maximize his expected payout, because R_p is strictly proper, and thus has no profitable deviation.

Since these assessments are both sequentially rational and consistent, they constitute a Perfect Bayesian Equilibrium of our game. Now, using a similar analysis, we can show that there is a Perfect Bayesian Equilibrium in our mechanism where all agents report truthfully; effectively, in the mechanism, every agent plays this game twice simultaneously, once as player 1 and once as player 2.

We can model our full mechanism as an extensive form game in a similar fashion, with simultaneous play in each round represented extensively with information sets. We claim that the following assessment by every agent i constitutes a Perfect Bayesian Equilibrium: play s_{z_i} upon receiving signal s_{z_i} from Nature in round 1, and play $\vec{p}_{\{s_{z_i}, x_h\}}$ after receiving signal x_h from player h in round 2, with the belief that you are at the node(s) in every information set where all previous players have played truthfully, and Nature assigned signals to players according to the common prior. The beliefs of all agents are clearly consistent with the strategy profiles of the agents, and there are no off-equilibrium paths. Furthermore, the assessments are sequentally rational, again because R_p is a strictly proper scoring rule. Consider the information set of any agent i in round 1; since he is only aware of his own signal, his belief about the distribution of s_{z_k}, the signal received by agent k, is $\vec{p}_{\{s_{z_i}\}}$. Conditional on agent j's signal s_{z_j}, his belief about the distribution of s_{z_k} is $\vec{p}_{\{s_{z_i},s_{z_j}\}}$; accordingly, he knows that $E(R_p(\vec{p}_{\{s_{z_i},s_{z_j}\}}, s_{z_k})) > E(R_p(\vec{p}', s_{z_k}))$ for $\vec{p}' \neq \vec{p}_{\{s_{z_i},s_{z_j}\}}$. However, since all other agents are playing truthfully, $x_k = s_{z_k}$, and thus $E(R_p(\vec{p}_{\{s_{z_i},s_{z_j}\}}, x_k)) > E(R_p(\vec{p}', x_k)$. Thus, agent i is uniquely maximizing his payout by playing $x_i = s_{z_i}$ in round 1, based on agent j's strategy profile, since that action maximizes the expected value of his information score, and his action in round one only influences his information score. Now consider any information set of an agent i in round 2; since he believes that all agents have played truthfully in round 1 - in particular agents h and j - his belief about the distribution of x_j is $\vec{p}_{\{s_{z_i},x_h\}}$, and thus he uniquely maximizes his prediction score by playing $y_i = \vec{p}_{\{s_{z_i},x_h\}}$.

Since the stated assessment is both sequentially rational and consistent, it is a Perfect Bayesian Equilibrium. Furthermore, since every agent's actions in the assessment uniquely maximizes his expected payout, this equilibrium is strict, as desired.

There is one essential assumption in the reasoning above, that agent i does not know x_k in round 1 and does not know x_j in round 2. The former is definitely true, since agent i has received no information other than his own signal in round 1. The latter is less clear, since in round 2, agent i knows x_h. However, for any $n \geq 3$ and any i, $h \neq j$. Thus, there is a Perfect Bayesian Equilibrium in our mechanism where all agents report truthfully for any $n \geq 3$. □

Theorem 3. *The sequential variant of the KFPP mechanism is ex-post individually rational for all admissible common priors and any $n \geq 3$ agents.*

Proof. Since this mechanism can use any strictly proper scoring rule, we can just choose a strictly proper scoring rule that is bounded below by zero; one choice would be an affine transformation of the quadratic rule, as mentioned in Chapter 3. Accordingly, all agents are guaranteed to have non-negative payout from this mechanism in all situations. □

4.2 Simultaneous Variant

4.2.1 Mechanism

For every agent i, select a reference agent j and a peer agent k. Every agent is asked for two reports:

- Information report: Let $x_i \in \{s_1, ...s_o\}$ be agent i's reported signal.
- Prediction report: Let $(\vec{y_i}^{s_1}, ..., \vec{y_i}^{s_o})$ be agent i's report about the frequencies of the signals conditional on j's signal, with $\vec{y_i}^{s_{z_j}}$ being agent i's predicted frequency vector, given that agent j received signal s_{z_j}.

Finally, agent i receives payout:

$$\underbrace{R_p(\vec{y_j}^{x_i}, x_k)}_{\text{Information Score}} + \underbrace{R_p(\vec{y_i}^{x_j}, x_k)}_{\text{Prediction Score}}$$

where R_p is any strictly proper scoring rule.

4.2.2 Equilibrium Analysis

The simultaneous variant of KFPP is very similar to the the sequential variant of KFPP. To avoid sequential interaction, agents do not actually receive the signal report of another agent. Instead, we virtualize this interaction by asking every agent to report every possible second-order posterior distribution conditional on another agent's signal. When we do finally receive another agent's signal, we resolve this vector of distributions to the actual second-order posterior distribution that the agent would have reported, had the mechanism been sequential. We will demonstrate that it is both Bayes-Nash incentive compatible and ex-post individually rational.

Theorem 4. *The simultaneous variant of the KFPP mechanism is strictly Bayes-Nash incentive compatible for all admissible common priors and any $n \geq 3$ agents.*

Proof. Fix some agent i, reference agent j, and peer agent k, with $i \neq j \neq k$. Assume that agent j and k report truthfully in both their information and prediction reports. We wish to show that the unique best response of agent i is to report truthfully. Notice that we can analyze agent i's information report (x_i) and prediction report $((\vec{y_i}^{s_1}, ..., \vec{y_i}^{s_o}))$ independently for best response criteria, because his total payout is a sum of his information and prediction scores, and his information score is dependent only on his information report and his prediction score is dependent only on his prediction report.

Information Report

We wish to determine the information report that maximizes agent i's information score. Denote agent j's realized signal by s_{z_j}; since we assumed that j is reporting truthfully, we have that his prediction report $(\vec{y_j}^{s_1}, ..., \vec{y_j}^{s_o}) = (\vec{p}_{\{s_{z_j},s_1\}}, ..., \vec{p}_{\{s_{z_j},s_o\}})$. Conditional on agent j's signal, agent i's true belief about the distribution of S_k is now $\vec{p}_{\{s_{z_j},s_{z_i}\}}$, where s_{z_i} denotes agent i's realized signal. Notice that submitting an information report of $x_i = s_l$ is equivalent to submitting the prediction

$$\vec{y_j}^{s_l} = \vec{p}_{\{s_{z_j},s_l\}}$$

on S_k to a strictly proper scoring rule, since agent i's information score is $R_p(\vec{y_j}^{x_i}, x_k)$, and we assume agent k is reporting truthfully, so x_k is the realized value of S_k. Finally, note that $\vec{y_j}^{s_l} = \vec{p}_{\{s_{z_j},s_l\}} \neq \vec{p}_{\{s_{z_j},s_m\}} = \vec{y_j}^{s_m}$ for $l \neq m$ because the common prior is admissible. Thus, since R_p is strictly proper, agent i uniquely maximizes his information score by submitting $x_i = s_{z_i}$, or reporting truthfully.

Prediction Report

We wish to determine the prediction report that maximizes agent i's prediction score. Denote agent j's realized signal by s_{z_j}; since we assumed that j is reporting truthfully, we have that his information report $x_j = s_{z_j}$. Conditional on this additional information, agent i's true belief about the distribution of S_k is now $\vec{p}_{\{s_{z_j},s_{z_i}\}}$, where s_{z_i} denotes agent i's realized signal. Notice that submitting a prediction report of $(\vec{y_i}^{s_1}, ..., \vec{y_i}^{s_o})$ is equivalent to submitting the prediction

$$\vec{y_i}^{s_{z_j}}$$

on S_k to a strictly proper scoring rule, since agent i's information score is $R_p(\vec{y_i}^{x_j}, x_k)$, and we assume agent k is reporting truthfully, so x_k is the realized value of S_k. Thus, agent i's expected prediction score is:

$$\sum_{a=1}^{o} P(s_{z_j} = s_a | S_i = s_{z_i}) \cdot \mathbb{E}[R_p(\vec{y_i}^{s_a}, x_k) | S_i = s_{z_i}, S_j = s_a]$$

Accordingly, since R_p is strictly proper, and agent i believes that $P(s_{z_j} = s_a | S_i = s_{z_i}) > 0$ for all $a \in \{1, ..., o\}$ (because the common prior is admissible), agent i uniquely maximizes his expected payout by setting

$$\vec{y_i}^{s_a} = \vec{p}_{\{s_a, s_{z_i}\}}$$

for all $a \in \{1, ..., o\}$; in other words, agent i uniquely maximizes his expected payout by reporting his prediction report truthfully.

Since agent i's unique best response is to report both his information report and prediction report truthfully, given that the other agents are also reporting truthfully, truthful report is a strict Bayes-Nash equilibrium. Thus, the KFPP is strictly Bayes-Nash incentive compatible for all admissible priors. □

Theorem 5. *The simultaneous variant of the KFPP mechanism is ex-post individually rational for all admissible common priors and any $n \geq 3$ agents.*

Proof. Since this mechanism can use any strictly proper scoring rule, we can just choose a strictly proper scoring rule that is bounded below by zero; one choice would be an affine transformation of the quadratic rule, as mentioned in Chapter 3. Accordingly, all agents are guaranteed to have non-negative payout from this mechanism in all situations. □

4.3 Tradeoffs

At first glance, the sequential variant of the KFPP mechanism seems strictly better than the simultaneous variant. While the former only requires that every agent report their signal and one probability distribution, the latter requires that every agent reports a vector of probability distributions in addition to their signal. This vector grows quadratically with the size of the signal space; for a signal space of size o, an agent will have to report o^2 individual probabilities under the simultaneous mechanism, but only o individual probabilities under the sequential mechanism. The complexity of the simultaneous mechanism makes it impractical for large signal spaces, as demanding such a large report from agents is impractical in real-world application. In contrast, the sequential variant of KFPP maintains a similar level of complexity to BTS and RBTS with respect to the reports demanded of the agent.

However, there is a significant drawback to the sequential mechanism: it requires all agents to report their signal before any agent reports his prediction. In many real-world applications, this is difficult to achieve, because agents participating in the mechanism often arrive at different times - for example, an online survey receives responses over time, rather than all at once - and it is unreasonable to expect an agent to wait until all agents have arrived to complete the mechanism. In practice, if one wishes to elicit signals from a total of n agents, one could conduct $\sim n/3$ iterations of the sequential mechanism, with each iteration of the mechanism being run after 3 agents have arrived; at least 3 agents are necessary because the sequential KFPP mechanism is not incentive compatible for fewer than 3 agents. However, one could imagine situations where the average wait time for just 3 agents to arrive is too high.

When this is the case, it may make sense to use the simultaneous KFPP mechanism, especially if the signal space is relatively small. Unlike an agent in the sequential KFPP, an agent in the simultaneous KFPP mechanism can give his information and prediction reports and receive a payout immediately; to compute his payout, the mechanism designer can randomly select two agents that arrived previously to be reference and peer agents. Only the first two agents to arrive must wait, since there are no previous agents with which to compute their payouts. Even then, they may make their information and prediction reports and leave, as long as they are willing to receive a delayed payout.

Chapter 5: Extensions

In this chapter, we consider several modifications to the base model presented in Chapter 3, and extend the KFPP mechanism to address those modifications. In section 5.1, we consider agents who are risk-adverse, rather than risk-neutral, and demonstrate that KFPP can easily handle risk-adversion with only slight modification. In section 5.2, we extend KFPP to handle continuous signal spaces. Finally, in section 5.3, we consider the possibility that agents incur a cost when acquiring and reporting a signal, and combine KFPP with a uniform auction to preserve incentive compatibility and individual rationality.

5.1 Risk Adversion

In our original model, described in Chapter 3, we assumed that all agents participating in the mechanism were risk-neutral - that their utility was linear with respect to the payout from the mechanism - and thus maximizing their expected utility is equivalent to maximizing their expected payout. However, in practice, most people have non-linear utility functions with respect to money; in particular, most people are risk-adverse, or have a concave down utility function. We consider the impact of agents with non-linear utility functions on KFPP, and propose two simple modifications to KFPP to handle them below; both are directly analogous to the methods proposed in Miller *et al.* (2005).

When agents have non-linear utility, KFPP will still be ex-post individually rational for such agents, since KFPP guarantees non-negative payout and thus non-negative utility. Unfortunately, KFPP's incentive compatibility may be threatened, because the action with the highest expected payout may no longer correspond to the action with the highest expected utility for an agent with a non-linear utility function.

If the mechanism designer knows the utility function U_i of each agent i, then preserving incentive compatibility is as simple as adjusting the payout of agent i from p to $U_i^{-1}(p)$. In this new mechanism, the utility of agent i is $U(U_i^{-1}(p)) = p$, the payout of agent i in the original mechanism. Consider the continuous KFPP. We showed that every agent uniquely maximized their expected payout by reporting their signal truthfully when all other agents were reporting their signal truthfully in the original mechanism. Accordingly, every agent must uniquely maximize their expected utility by reporting their signal truthfully when all other agents are reporting their signal truthfully in this modified mechanism. However, this implies that the modified mechanism is Bayes-Nash incentive compatible. Similar logic applies for the simultaneous KFPP.

However, if the mechanism designer does not know the utility functions of the agents, as is often the case, we can utilize a property of all valid Von Neumann-Morgentern utilities to preserve incentive compatibility. Von Neumann utilities are all linear with respect to probabilities, so if we replace the payouts in our original mechanism with lottery tickets to a binary-outcome lottery, all agents will maximize their expected utility by maximizing expected payout from the mechanism. Intuitively, regardless of the shape of their utility function, all agents will want to maximize the probability that they win the lottery (since their utility must be monotone with respect to money), and they do this by maximizing the expected number of lottery tickets they receive.

Accordingly, KFPP has no difficulty in handling agents with varying risk-preferences.

5.2 Continuous Signals

Most work on eliciting truthful reports of private signals when objective truth is inaccessible has focused on discrete signal spaces. As a practical matter, this is reasonable; most proposed mechanisms require agents to report a distribution on the signal space, which might be too onerous (or literally impossible due to lack of expressiveness in the mechanism) for the average agent when the signal space is continuous. However, when the mechanism designer is eliciting information from a group of experts about a naturally continuous signal, a mechanism that can handle a continuous signal space may be useful;

regardless, we believe it to be of theoretical interest if nothing else.

First, we review previous mechanisms (enumerated in the related works section) with respect to their ability to accomodate a continuous signal distribution. As we mentioned previously, the shadowing technique used in the RBTS mechanism doesn't even extend nicely to a n element signal space, much less a continuous signal space. In contrast, the vanilla BTS asks agents to answer an m multiple-choice question for any finite m, and thus is suitable for an arbitrary discrete and finite signal space. Furthermore, with slight modification in framing, BTS can handle a discrete signal space that is countably infinite. However, the general technique employed by BTS - the surprisingly common criterion - does not have a natural analogue for continuous signal spaces. In particular, given any finite number of participants, the mechanism designer in BTS will calculate a discrete probability distribution for the population endorsement frequencies, and a continuous probability distribution for the predicted frequencies. Accordingly, the mechanism designer will not be able to calculate either an information score or a prediction score, which depend on both the population endorsement frequencies and predicted frequencies, because of this inconsistency. In short, it is no long clear what is meant by "surprisingly common," because any particular realized signal has probability zero under a continuous probability distribution. Finally, the PP method can handle a continuous signal space, by substituting a continuous strictly proper scoring rule for the traditional discrete strictly proper scoring rule used in the mechanism.

Due to the similarity between KFPP and the PP method, the same technique in Miller *et al.* (2005) can be utilized by KFPP to accomodate continuous signal spaces. Specifically, since KFPP depends solely on scoring rules in the calculation of payouts, one can easily substitute a continuous strictly proper scoring rule for the discrete strictly proper scoring rule used in KFPP. Accordingly, KFPP retains all of the expressive power of the Peer Prediction mechanism, while removing the requirement that the mechanism designer knows the common prior - a strong assumption that restricts the applicability of the mechanism - and thus improves on RBTS in expressiveness, and BTS in both robustness and expressiveness.

5.3 Effort Elicitation

We have shown that KFPP is incentive compatible and ex-post individually rational for agents who do not incur any cost during the mechanism. However, in practice, agents who participate in the mechanism may have non-zero costs associated with reporting a truthful signal. We now consider the impact of these costs on KFPP.

We will model costs as follows. Each agent i has some private type $c_i > 0$, which represents the fixed cost associated with reporting his signal truthfully, and the c_i are distributed according to some distribution C. Note that we intentionally make no assumptions about when this cost occurs, just that the cost is definitely incurred when an agent reports his signal truthfully, and can be avoided by the agent; for example, this cost maybe associated with entering the mechanism, acquiring the signal, or reporting the signal. Our goal is to construct some mechanism M that is both ex-post individually rational, and guarantees that all agents acquire and report a signal truthfully, regardless of their type.

If C is bounded above, and the mechanism designer knows this bound, the mechanism designer may be able to preserve both incentive compatibility and ex-post individual rationality by adding $\max(C)$ to all payments made by KFPP, depending on when during the mechanism the cost is incurred. However, if either one of these conditions is not satisified, we will show that it is not possible to construct such a mechanism. In fact, we cannot construct a mechanism that is incentive compatible and interim individually rational; not all agents will have a positive expected payout after learning their type/private cost.

Consider any mechanism M with strategy space $S = S_1 \times S_2 \times ... \times S_n$, where S_i is the strategy space for agent i, outcome space O, and outcome rule $F : S \to \Pi(O)$, where $\Pi(O)$ denotes the set of distributions over O. We make two simplifying assumptions. First, in our context, agents are not intrinsically interested in the outcome of the mechanism, aside from any payments made by the mechansim to the agents, and thus we can write $O = \mathbb{R}^n$, where for any $o = (o_1, ..., o_n) \in O$, o_i is the payment made by the mechanism to agent i. Second, we require that for any i, S_i can be partitioned into two non-empty sets C_i, where agent i is guaranteed to incur cost c_i, and C_i^c where agent i does not incur cost

c_i; furthermore, there must be a non-empty subset of C_i where agent i is guaranteed to acquire and report a signal truthfully. This second assumption just guarantees that M actually solicits signals from the agents.

Theorem 6. *If C is not bounded above, then M does not have an interim individually rational Bayes-Nash equilibrium where all agents are guaranteed to report their signal truthfully.*

Proof. Consider any Bayes-Nash equilibrium $s^* \in S$ in M where all agents are guaranteed to report their signal truthfully. We apply the *Revelation Principle* to arrive at some direct-revelation mechanism M' with a payoff-equivalent Bayes-Nash equilibrium where all agents report their type truthfully. In particular, assuming that the agents can yield control over their choice to acquire and report a signal truthfully (or not) to the mechanism, the mechanism $M' = (S', F')$, where $S' = (\mathbb{R}^+)^n$ and $F' = F \circ s^*$, has a Bayes-Nash equilibrium where every agent reports their type truthfully, and the payout of this mechanism is the same as the original mechanism. In M', an agent's only action is to report their type (not necessarily truthfully), and M' simulates the original mechanism with the optimal strategies in the original Bayes-Nash equilibrium. We claim that truthful reporting is a Bayes-Nash equilibrium in M'. Assume to the contrary, that some agent i can profitably deviate in M' when her type $c_i = c'$ by reporting $c'' \neq c'$ when all other agents are reporting truthfully. Then s^* could not have been a Bayes-Nash equilibrium in M, because agent i could have profitably deviated by playing

$$s'_i(c_i) = \begin{cases} s^*_i(c'') \text{ if } c_i = c' \\ s^*_i(c_i) \text{ otherwise} \end{cases}$$

in M; this contradicts our original assumption about M, and thus there is a Bayes-Nash equilibrium in M' where all agents report their type truthfully. To see why this equilibrium in M' is payoff-equivalent to M, note that $F'(c) = F(s^*(c))$, where c is the true type profile of the agents; since all agents are reporting their true type in this equilibrium, the payout profile for this equilibrium in M' must be equal to the payout profile for the original equilibrium in M.

We define the function $p_i : O \to \mathbb{R}$ such that for any $o \in O$, $p_i(o)$ is the payment made to agent i in outcome o. Consider the situation where agent 1 has type k for some fixed constant $k > 0$. Let $t = \mathbb{E}[p_1(F'(k, c_{-1}))|C]$ be agent 1's expected payout from M' when he has type k. Now consider the situation where agent 1 has type $k' = k + t$, and let $t' = \mathbb{E}[p_1(F'(k', c_{-1}))|C]$ be agent 1's expected payout from M' when he has type k'. We claim that $t \geq t'$. To see why this must be the case, assume to the contrary, that $t < t'$; then truthful reporting can not be a Bayes-Nash equilibrium, since when agent 1 has type k he could profitably deviate by reporting k' instead, if all other agents are reporting truthfully. Accordingly, the equilibrium in M' where all agents report their type truthfully is not interim individually rational for agent 1, since all agents are guaranteed to acquire and report their signal under s^*; thus when agent 1 has type k', he incurs cost k' and receives an expected payout of $t' - k' = t' - (k + t) \leq t - (k + t) = -k < 0$ from M'.

However, since the truthful equilibrium in M' is payoff equivalent to our original equilibrium in M, our original equilibrium could not have been interim individually rational for agent 1 either. Since this is true for any Bayes-Nash equilibrium in M where all agents are guaranteed to report their signal truthfully, there must be no Bayes-Nash equilibrium in M that is both interim individually rational, and guarantees that all agents acquire a signal and report it truthfully, as desired. □

Theorem 7. *If C is not known to the mechanism designer, then M does not have an interim individually rational Bayes-Nash equilibrium where all agents are guaranteed to report their signal truthfully.*

Proof. We can use similar reasoning to the previous proof to arrive at this result. □

Note that the two previous theorems do not preclude the possibility that for any fixed constant $k > 0$, an agent with type k will find at least one interim individually rational equilibrium in M where all agents are reporting their signal truthfully; this may occur if M has an infinite number of equilibria. However, mechanisms with this property are not desirable, because it is unlikely that the agents will converge to the desired equilibrium, for two reasons. First, an infinite number of equilibria naturally makes equilibrium selection difficult. Second, since any particular agent does not know the types of other agents, he cannot identify the equilibria that are interim individually rational for other agents, making convergence even less likely.

These impossibility result are unfortunate; it means that we cannot, in general, elicit truthful signals from all agents if they experience unknown fixed cost when reporting their signal. However, assuming that agents incur their fixed cost when acquiring a signal, and that the mechanism designer can observe/control when an agent acquires a signal, we can construct a mechanism that is both interim individually rational, and has an equilibrium where all but one of the agents acquire a signal and report it truthfully. While these assumptions certainly limit the applicability of the following mechanism, they are reasonable in a wide variety of real-world situations, including when the mechanism designer also controls the distribution of signals - for example, when Amazon wants customer feedback on products it sells - or when soliciting the signal immediately causes the agent to acquire the "signal" - for example, when a survey question about a personal opinion immediately causes the responder to reflect on the question.

To construct our mechanism, we will combine a uniform auction with the original KFPP mechanism. Informally, we allow all agents in the mechanism to bid on a seat in the original KFPP mechanism; the agent that submits the worst bid (the highest cost for participating in the KFPP mechanism) does not get to participate in the original KFPP mechanism, but all other agents do, and receive a payment of the worst bid in addition to any payment from the KFPP mechanism. A more formal description of the mechanism, which we will denote by M, can be found below:

1. Step 1: All players report their cost c_i for acquiring a signal (not necessarily truthfully).

2. Step 2: Let $m = \text{argmax}_i c_i$. Player m receives payment 0, and exits the mechanism. All other players acquire a signal and incur any cost associated with this acquisition; we will renumber the remaining agents $1, ..., n-1$ for convenience.

3. Step 3: All remaining players simultaneously report their signal $x_i \in \{s_1, ..., s_o\}$ to the mechanism.

4. Step 4: Every remaining player i receives the report of player $i-1$, x_i, from the

mechanism, and then reports the frequency of the signals $\vec{y_i} = (y_i^1, ..., y_i^o)$.

5. Step 5: Every remaining player i receives payment $R_p(\vec{y_j}, x_k) + R_p(\vec{y_i}, x_j) + c_i$, where R_p is a strictly proper scoring rule bounded below by zero.

Theorem 8. *M is interim individually rational, and has a strict Perfect Bayesian equilibrium where all but one agent acquire and report their signals truthfully.*

Proof. First, we note that for the subgame starting at step 3, it is a strict Perfect Bayesian equilibrium for all remaining agents to report their signal truthfully, because of our previous analysis of the KFPP mechanism; this is due to the fact that linear transformations of strictly proper scoring rules are still strictly proper, so the incentive structure of the mechanism has not changed. Now, we will argue that it is a dominant strategy for all agents to report $c_i - \mathbb{E}[R_p(\vec{y_j}, x_k) + R_p(\vec{y_i}, x_j)|p]$, where p is the common prior over the signals, in step 1 of this mechanism, assuming the beliefs we derived for KFPP earlier in the subgame starting at step 3. For convenience, we will denote $k = \mathbb{E}[R_p(\vec{y_j}, x_k) + R_p(\vec{y_i}, x_j)|p]$; note that k is the expected payout of the KFPP mechanism with the belief that everyone will report their signal and signal distributions truthfully and that signals are distributed according to p, and thus bidding $c_i - k$ is bidding the agent's true value for the KFPP game. Let c_i and b_i be the type and bid of agent i, respectively. The expected payoff of agent i is:

$$\begin{cases} \max_{j \neq i} b_j + k - c_i \text{ if } b_i < \max_{j \neq i} b_j \\ 0 \text{ otherwise} \end{cases}$$

First, we show that bidding $b_i > c_i - k$ is dominated by bidding truthfully. If $c_i - k < b_i < \max_{j \neq i} b_j$, then agent i would have received the same payout $\max_{j \neq i} b_j + k - c_i$ by bidding truthfully as bidding b_i. If $c_i - k < \max_{j \neq i} b_j < b_i$ then agent i would have received a higher payout $\max_{j \neq i} b_j + k - c_i > 0$ by bidding truthfully than by bidding b_i. Finally, if $\max_{j \neq i} b_j < c_i - k < b_i$, then agent i would have received the same payout 0 by bidding truthfully as bidding b_i.

Similarly, bidding $b_i < c_i - k$ is dominated by bidding truthfully. If $b_i < c_i - k < \max_{j \neq i} b_j$, then agent i would have received the same payout $\max_{j \neq i} b_j + k - c_i$ by bidding truthfully as bidding b_i. If $b_i < \max_{j \neq i} b_j < c_i - k$ then agent i would have received a higher payout $0 > \max_{j \neq i} b_j + k - c_i$ by bidding truthfully than by bidding b_i. Finally, if $\max_{j \neq i} b_j < b_i < c_i - k$, then agent i would have received the same payout 0 by bidding truthfully as bidding b_i.

Since bidding $c_i - k$ is a dominant strategy in step 1, the assessment where every agent reports $c_i - k$ in step 1, reports their signal and signal distribution truthfully in steps 3-5, and believes that all other agents who make it to step 3 also report truthfully afterwards, is a strict Perfect Bayesian equilibrium; this follows from our analysis above and our analysis of the KFPP mechanism earlier. Note that the agents can have any belief about the behavior of the other agents in step 1, because their choice of action is dominant.

Furthermore, this Perfect Bayesian equilibrium has all the properties we desired. By construction, $n - 1$ agents report their signal truthfully in this equilibrium. In addition, the expected payout to every agent in this equilibrium is non-negative after the agents have received their types, because agents will either receive payout 0 if they do not make it past step 2, or receive an expected positive payout if they do make it past step 2. □

The above mechanism used the sequential variant of KFPP, but a similar mechanism can be constructed with the simultaneous variant of KFPP as well. We make three additional observations about the properties of this mechanism.

First, we've only shown that the mechanism is interim individually rational, so the mechanism may have negative payouts to agents in some situations. While we previously said that this was undesirable, we argue that it is acceptable in this situation. If the cost associated with acquiring and reporting a signal is not paid to the mechanism designer - for example a cost that is internal to the agents, like an effort cost associated with acquiring a signal - then the mechanism designer never demands transfers from the agents even when the mechanism results in a negative payout; the only transfers that occur between the mechanism and the agents are dictated by KFPP, which is guaranteed to have non-negative payout by itself. In situations where the cost associated with acquiring and reporting the signal is paid to the mechanism designer - for example, the purchase of a product to be rated from Amazon - then the mechanism designer generally has no practical difficulty with demanding payments from the agents anyway.

Second, this mechanism can be adjusted to elicit any desired number of signals n' for

$1 < n' < n$, by allowing only the n' agents with the lowest reported cost to proceed past step 2, instead of $n - 1$ agents. In this way, a mechanism designer who is interested in eliciting the information of a specific fraction of the population can do so without difficulty.

Third, every agent reports their true expected value for playing the KFPP mechanism in M; their expected value for KFPP is their true cost/type minus a constant (the expected value for playing KFPP for a costless agent). Accordingly, in practice the mechanism designer can learn about the cost distribution of the agents through this mechanism, which can help with running future iterations of the mechanism, or may be intrinsically interesting to the mechanism designer.

Chapter 6: Conclusion

In this thesis, we have presented Knowledge Free Peer Prediction, a mechanism designed to elicit private information from individuals when objective truth is inaccessible. It overcomes several difficulties that previous proposed mechanisms - like Peer Prediction, the Bayesian Truth Serum, and the Robust Bayesian Truth Serum - experience; not only can the mechanism be run by a mechanism designer who is ignorant of the common prior beliefs of the participants, the mechanism is incentive compatible and ex-post individually rational given at least 3 participants. To do this, it mimics Peer Prediction, but delegates any Bayesian updating normally performed by the mechanism designer to the participants of the mechanism. Furthermore, while we constructed KFPP in an ideal setting, we demonstrated that KFPP can be modified to handle many real-world challenges, including risk-adverse agents, continuous signals, and effort elicitation from agents who experience costs. Accordingly, it is suitable for application in a wide variety of situations.

6.1 Future Work

There are several avenues available for future investigation.

First, the assumption that all participants share a common prior is likely unrealistic in real-world settings. Witkowski and Parkes (2012a) proposes two related mechanisms that successfully elicit private information even when participants have different prior beliefs. However, both mechanisms assume that the signal space is binary, much like the Robust Bayesian Truth Serum. Accordingly, one area for future work would be the extension of a mechanism like KFPP, that can handle all discrete and continuous signal spaces, to situations where no common prior exists.

Second, experimental verification of the KFPP mechanism would be highly beneficial. We do not expect actual participants of the KFPP mechanism to explicitly perform Bayesian updating, nor consider the equilibrium analysis of KFPP. Accordingly, for KFPP to be practically applicable as well as theoretically interesting, experimental verification of the truthfulness of the mechanism, and experimental comparison of KFPP with other mechanisms like BTS and RBTS, is necessary.

Finally, further work on the truthful elicitation of continuous signals is necessary. It would be particularly interesting to see a mechanism that can induce truthful reporting when the question being asked naturally demands an open-form answer. While KFPP and Peer Prediction can both theoretically handle such continuous signals, it's unclear how to formulate and report a distribution on such a signal space. Adapting information elicitation mechanisms to this domain would allow for the elicitation of much more complicated and interesting information.

Bibliography

G. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

A. Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

V. Krishna. *Auction theory.* Academic Press, San Diego, California, 2002.

J. Matheson and R. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1086–1096, 1976.

N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

R. Myerson. Incentive compatibility and the bargaining problem. *Econometrica*, 47(1):61–73, 1979.

D. Prelec. A bayesian truth serum for subjective data. *Science*, 306:462–466, 2004.

R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1:43–62, 1998.

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

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

Appendix A: Appendix

A.1 Appendix 1

While the natural extension of the Robust Bayesian Truth Serum is not strictly Bayes-Nash incentive compatible for all admissible common priors, it turns out that it works for many reasonable common priors. We characterize one such class of common priors below.

Theorem 9. *The natural extension of the Robust Bayesian Truth Serum is strictly Bayes-Nash incentive compatible for any number of agents $n \geq 3$ and all admissible priors with the property that:*

$$p^{s_a}_{\{s_a,s_b\}} - p^{s_a}_{\{s_b\}} > p^{s_c}_{\{s_a,s_b\}} - p^{s_c}_{\{s_b\}}$$

for any $a, b, c \in \{1, ..., o\}$ where $a \neq c$.

Proof. Fix some agent i, reference agent j, and peer agent k. Assume that agent j and k report truthfully in both their information and prediction reports. We wish to show that the unique best response of agent i is to report truthfully. Notice that we can analyze agent i's information report (x_i) and prediction report (y_i) independently for best response criteria, because her total payout is a sum of her information and prediction scores, and her information score is dependent only on her information report and her prediction score is dependent only on her prediction report.

Information Report

We wish to determine the information report that maximizes agent i's information score. Denote agent j's realized signal by s_{z_j}; since we assumed that j is reporting truthfully, we have that $\vec{y_j} = \vec{p}_{\{s_{z_j}\}}$. Conditional on agent j's signal, agent i's true belief about the distribution of S_k is now $\vec{p}_{\{s_{z_j},s_{z_i}\}}$, where s_{z_i} denotes agent i's realized signal. Notice that submitting an information report of $x_i = s_l$ is equivalent to submitting the prediction

$$\vec{y_i'} = \vec{y_j} + \vec{\delta_l} = \vec{p}_{\{z_j\}} + \vec{\delta_l}$$

on S_k to the quadratic scoring rule, since agent i's information score is $R_q(y_i', x_k)$, and we assume agent k is reporting truthfully, so x_k is the realized value of S_k. By Lemma 1, agent i's best response is thus to report $x_i = s_l$ such that

$$||\vec{p}_{\{s_{z_j},s_{z_i}\}} - (\vec{p}_{\{s_{z_j}\}} + \vec{\delta_l})||^2$$

is minimized. We claim that submitting $x_i = s_{z_i}$ (reporting her belief) uniquely minimizes this distance. To see this, consider any alternative report s_l such that $s_l \neq s_{z_i}$.

$$\begin{aligned}
&||\vec{p}_{\{s_{z_j},s_{z_i}\}} - (\vec{p}_{\{s_{z_j}\}} + \vec{\delta_{z_i}})||^2 - ||\vec{p}_{\{s_{z_j},s_{z_i}\}} - (\vec{p}_{\{s_{z_j}\}} + \vec{\delta_l})||^2 \\
=&\sum_{k=1,k\neq z_i}^{o} (p^{s_k}_{\{s_{z_j},s_{z_i}\}} - (p^{s_k}_{\{s_{z_j}\}} - \frac{\delta}{(o-1)}))^2 + (p^{s_{z_i}}_{\{s_{z_j},s_{z_i}\}} - (p^{s_{z_i}}_{\{s_{z_j}\}} + \delta))^2 \\
&- \sum_{k=1,k\neq l}^{o} (p^{s_k}_{\{s_{z_j},s_{z_i}\}} - (p^{s_k}_{\{s_{z_j}\}} - \frac{\delta}{(o-1)}))^2 - (p^{s_l}_{\{s_{z_j},s_{z_i}\}} - (p^{s_l}_{\{s_{z_j}\}} + \delta))^2 \\
=&(p^{s_{z_i}}_{\{s_{z_j},s_{z_i}\}} - (p^{s_{z_i}}_{\{s_{z_j}\}} + \delta))^2 - (p^{s_{z_i}}_{\{s_{z_j},s_{z_i}\}} - (p^{s_{z_i}}_{\{s_{z_i}\}} - \frac{\delta}{(o-1)}))^2 \\
&+ (p^{s_l}_{\{s_{z_j},s_{z_i}\}} - (p^{s_l}_{\{s_{z_j}\}} - \frac{\delta}{(o-1)}))^2 - (p^{s_l}_{\{s_{z_j},s_{z_i}\}} - (p^{s_l}_{\{s_{z_i}\}} + \delta))^2
\end{aligned}$$

To simplify notation, we will substitute $a = p^{s_{z_i}}_{\{s_{z_i},s_{z_i}\}} - p^{s_{z_i}}_{\{s_{z_j}\}}$ and $b = p^{s_l}_{\{s_{z_j},s_{z_i}\}} - p^{s_l}_{\{s_{z_j}\}}$.

$$\begin{aligned}
=&(a-\delta)^2 - (a + \frac{\delta}{(o-1)})^2 + (b + \frac{\delta}{(o-1)})^2 - (b-\delta)^2 \\
=&a^2 + \delta^2 - 2a\delta - a^2 - \frac{\delta^2}{(o-1)^2} - 2a\frac{\delta}{(o-1)} \\
&+ b^2 + \frac{\delta^2}{(o-1)^2} + 2b\frac{\delta}{(o-1)} - b^2 - \delta^2 + 2b\delta \\
=&2\delta(b-a) + \frac{2\delta}{o-1}(b-a)
\end{aligned}$$

However, by assumption, we have that $b = p^{s_l}_{\{s_{z_j}, s_{z_i}\}} - p^{s_l}_{\{s_{z_j}\}} < p^{s_{z_i}}_{\{s_{z_i}, s_{z_i}\}} - p^{s_{z_i}}_{\{s_{z_j}\}} = a$. Since $\delta = \min(\min_i(y^i_j), \min_i(1 - y^1_j)) = \min(\min_i(p^{s_i}_{z_j}), \min_i(1 - p^{s_i}_{z_j})) > 0$ (because the common prior is admissible, and thus fully mixed), we have that

$$\begin{aligned}
&||\vec{p}_{\{s_{z_j}, s_{z_i}\}} - (\vec{p}_{\{s_{z_j}\}} + \vec{\delta_{z_i}})||^2 - ||\vec{p}_{\{s_{z_j}, s_{z_i}\}} - (\vec{p}_{\{s_{z_j}\}} + \vec{\delta_l})||^2 \\
=&2\delta(b - a) + \frac{2\delta}{o - 1}(b - a) \\
<&0
\end{aligned}$$

Accordingly, agent i's best response is to report truthfully with respect to her information report.

Prediction Report

Assuming that agent k reports truthfully, x_k is the realized value of S_k. Accordingly, by the strict properness of the quadratic scoring rule, the unique report that maximizes agent i's prediction report

$$R_q(y_i, x_k)$$

is to submit $y_i = \vec{p}_{\{s_{z_i}\}}$, where s_{z_i} is the signal received by agent i.

Since agent i's unique best response is to report both her information report and prediction report truthfully, given that the other agents are also reporting truthfully, truthful report is a strict Bayes-Nash equilibrium. Thus, the Robust Bayesian Truth Serum is strictly Bayes-Nash incentive compatible for all admissible priors with the property that

$$p^{s_a}_{\{s_a, s_b\}} - p^{s_a}_{\{s_b\}} > p^{s_c}_{\{s_a, s_b\}} - p^{s_c}_{\{s_b\}}$$

for any $a, b, c \in \{1, ..., o\}$ where $a \neq c$. □

Note that the condition provided above is sufficient, but not necessary; there are other common priors not captured above for which the natural extension to the Robust Bayesian Truth Serum is strictly Bayes-Nash incentive compatible. Furthermore, even this class of common priors is arguably relatively large, in the sense that it is likely that many real-world common priors, if they exist, would satisfy the constraint that $p^{s_a}_{\{s_a, s_b\}} - p^{s_a}_{\{s_b\}} > p^{s_c}_{\{s_a, s_b\}} - p^{s_c}_{\{s_b\}}$ for any $a, b, c \in \{1, ..., o\}$ where $a \neq c$. In situations where this constraint is not satisfied - when two signals s_a and s_c are strongly correlated, and thus $p^{s_a}_{\{s_a, s_b\}} - p^{s_a}_{\{s_b\}} \leq p^{s_c}_{\{s_a, s_b\}} - p^{s_c}_{\{s_b\}}$ - we might suggest grouping the correlated signals together as a practical solution to make the mechanism incentive compatible.

More generally, given some coarse knowledge about the common prior, the mechanism designer could modify the manner in which the mechanism "shadows" (modify $\vec{\delta_i}$) to make the mechanism incentive compatible.