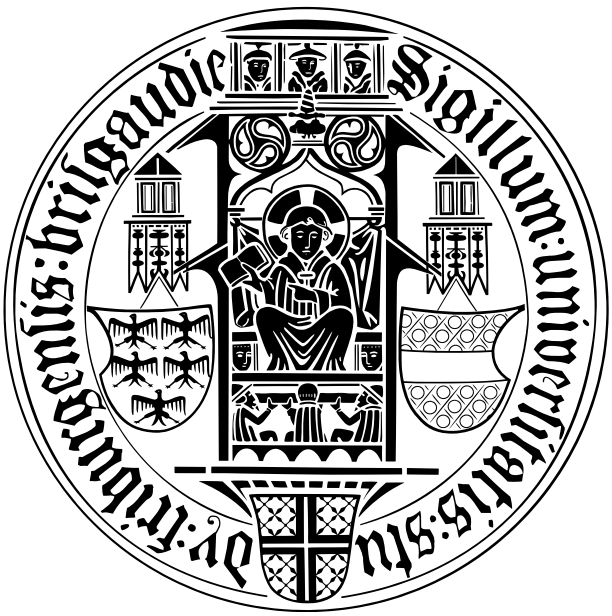
Dissertation zur Erlangung des Doktorgrades der Technischen Fakultät
der Albert-Ludwigs-Universität Freiburg im Breisgau

Robust Peer Prediction Mechanisms

Vorgelegt von
Dipl.-Inf. Jens Witkowski

am 21.04.2014



Betreut von
Prof. Dr. Bernhard Nebel
Prof. David C. Parkes, Ph.D.

Tag der Disputation:	05.05.2014
Dekan:	Prof. Dr. Georg Lausen
Vorsitz:	Prof. Dr. Wolfram Burgard
Beisitz:	Prof. Dr. Peter Fischer
Betreuer:	Prof. Dr. Bernhard Nebel
Prüfer:	Prof. David C. Parkes, Ph.D.

Game theory has a great advantage in explicitly analyzing the consequences of trading rules that presumably are really common knowledge; it is deficient to the extent it assumes other features to be common knowledge, such as one player's probability assessment about another's preferences or information.

I foresee the progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analyses of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality.

Wilson [1987]

Acknowledgements

First and foremost, I am deeply thankful for my two outstanding advisors, Prof. Bernhard Nebel from Albert-Ludwigs-Universität Freiburg and Prof. David Parkes from Harvard University. I thank Bernhard for providing me with an environment in which I could grow as a researcher and for always supporting my career endeavors. I particular, I am deeply grateful that, right from the beginning, Bernhard gave me the academic freedom to develop my own research agenda which now lead to this thesis.

In my second year, I moved to Cambridge, MA to begin what I thought would be a four month research visit of the EconCS group at Harvard. The main reason for why this visit turned out to be longer is David, whom I consider an academic superhero. There is much to be said about his distinction as a scientist, and the awe after receiving detailed comments on a 15-page paper you sent him 10 minutes ago. Most importantly though, David is a truly exceptional mentor with a wonderful sense for people. I am deeply grateful for his contagious excitement for research and his kind support of my work and ideas.

I also thank Prof. Malte Helmert and Prof. Sven Seuken. Malte helped me with my initial steps into research during his time as a postdoc in Bernhard's group, and Sven originally introduced me to David and has been a close friend for a long time. In addition to Sven's primary role as a friend, I am grateful to have him as collaborator, travel partner, fellow froyo aficionado, and rational agent (the non-selfish kind).

I thank the Landesgraduiertenförderung Baden-Württemberg and the German Academic Exchange Service (DAAD) for their support through their PhD fellowships. I am also grateful to Yoram Bachrach and Peter Key for giving me the opportunity to intern with them at Microsoft Research in Cambridge (UK).

I thank the members of the Foundations of Artificial Intelligence and EconCS groups for many interesting discussions. A special thanks goes to Rafael Frongillo, Malvika Rao, Sven Seuken, and Bo Waggoner for proofreading this thesis and for their helpful feedback, and to Robert Mattmüller for logistical help.

I thank my family and friends for their unconditional love and support.

Abstract

This thesis addresses the challenge of peer prediction, which seeks to elicit private information from rational agents without the requirement that ground truth is eventually revealed. The classical peer prediction method provides a solution to the peer prediction challenge. It compares the reported information of an agent with the reported information of another agent, and computes a payment rule that implements truth revelation in a strategic equilibrium. However, the algorithm computing the payments critically depends on the method's assumption that all agents share the same prior beliefs and that the algorithm ("mechanism") knows these beliefs.

In this thesis, I relax this common knowledge assumption. I first design the *Robust Bayesian Truth Serum* (RBTS), which asks agents for two types of reports, the report of the private information it is interested in and a prediction report corresponding to an agent's belief about the private information of other agents. RBTS dispenses with the assumption that the agents' prior beliefs need to be known to compute the payments. It does, however, still rely on the agents sharing the same prior beliefs. My second contribution is the design of *subjective-prior peer prediction mechanisms*, which further reduce the assumption of common knowledge. As in RBTS, they do not require knowledge of the agents' prior beliefs. Moreover, they allow the agents' prior beliefs to be subjective, i.e. different from one another. My third contribution is the study of *effort-incentivizing peer prediction*. In many applications of interest, the information that seeks to be elicited first needs to be acquired. When this information acquisition requires costly effort, agents may have an incentive to avoid it and choose to guess instead. Addressing this problem, I suggest payments, where only agents with good enough information have an incentive to participate in the mechanism. Agents not investing effort and agents with low quality, choose to pass, effectively self-selecting according to quality.

Zusammenfassung

Diese Dissertation beschäftigt sich mit Peer-Prediction-Mechanismen, welche Nutzern von Online-Systemen einen Anreiz geben, private Meinungen oder Erfahrungen ehrlich abzugeben. Der ursprüngliche Peer-Prediction-Mechanismus [Miller et al., 2005] vergleicht dazu die Antworten von zwei Nutzern miteinander und bezahlt diese anhand einer Zahlungsregel, welche sicherstellt, dass die Übermittlung von ehrlichen Informationen ein spieltheoretisches Gleichgewicht bildet. Das Problem dieses Mechanismus ist, dass er für die Praxis zu hohe Anforderungen an das gemeinsame Wissen (*common knowledge*) der Nutzer stellt.

In dieser Dissertation entwickle ich Peer-Prediction-Mechanismen mit stark abgeschwächten Annahmen an das gemeinsame Wissen, was eine praktische Anwendung dieser Mechanismen ermöglicht. Ich entwickle zuerst das *Robust Bayesian Truth Serum* (RBTS), welches, zusätzlich zur Meinung oder Erfahrung des Nutzers, auch eine probabilistische Vorhersage der übermittelten Erfahrungen anderer Nutzer abfragt. Im Gegensatz zum Original-BTS-Mechanismus ist RBTS bereits ab drei Nutzern anreizkompatibel und kann außerdem so konfiguriert werden, dass die Zahlungen an die Nutzer nie negativ sind. Aufbauend auf RBTS schwäche ich die Anforderungen an gemeinsames Wissen weiter ab und entwickle einen anreizkompatiblen Mechanismus, in dem jeder Nutzer nicht nur subjektive Erfahrungen macht, sondern auch ein subjektives Modell der Wirklichkeit hat. Wie bei RBTS müssen die Nutzer sowohl die eigentliche Information als auch eine Vorhersage über die von anderen Nutzern abgegebenen Informationen übermitteln. Im Unterschied zu RBTS kann der Mechanismus dafür jedoch nicht auf die von anderen Nutzern übermittelten probabilistischen Vorhersagen zurückgreifen. Während alle diese Mechanismen den Nutzern Anreize geben, Informationen ehrlich zu übermitteln, müssen Nutzer diese Informationen in vielen Anwendungen, insbesondere im *crowdsourcing*, jedoch erst beschaffen. Ich entwickle hierfür einen Mechanismus, der den Nutzern einen Anreiz gibt, den für die Informationsbeschaffung nötigen Aufwand zu betreiben. Nutzer, die sich bei einer Antwort nicht sicher sind, gibt der Mechanismus darüber hinaus einen Anreiz zu passen und keine Bewertung abzugeben.

Contents

Chapter 1

Introduction

1.1 Peer Prediction

User-generated content is essential to the effective functioning of many social computing and e-commerce platforms. Prominent examples include the elicitation of feedback about products or services on sites such as Amazon Reviews[1] or Expedia[2], and the elicitation of information from workers on crowdsourcing platforms, such as Amazon Mechanical Turk[3]. On these platforms, workers are paid small rewards to do so-called *human computation* tasks, which are tasks that are easy to solve for humans but difficult for computers. For example, humans have no problem recognizing a celebrity in an image they are shown, whereas even state-of-the-art computer vision algorithms are still not capable of solving this task with sufficient accuracy. While statistical estimation techniques can be adopted for the purpose of adjusting for biases or identifying users whose inputs are especially noisy, they are appropriate only in settings with repeated participation by the same user and when user inputs are informative in the first place. But what if providing accurate information is costly for users, or if users have incentives to submit false reports?

Consider a worker in a crowdsourcing context who labels images, so that they can be indexed by search engines, and who would rather avoid investing effort by typing in random words as labels or choosing labels that are too generic (e.g. "man"). Alternatively, consider the crowdsourcing task of labeling websites that contain inappropriate content for an advertiser. When workers are paid a fixed amount per task, they can improve their hourly rate by skipping over tasks without investing due effort. Or consider a web service that wishes to

[1] `http://www.amazon.com`
[2] `http://www.expedia.com`
[3] `http://www.mturk.com`

publish the expected download speed of a file mirrored on different server sites. A user who found a fast server may be concerned that sharing truthful, positive feedback could cause the server to become more popular, slowing down future downloads. Moreover, there are settings where privacy is a concern. Consider a public health program that requires participants to report whether they have ever used illegal drugs, and where participants may lie about their drug abuse because of shame or concerns about not being eligible for the program.

Peer prediction mechanisms address these incentive problems. They are designed to elicit truthful private information from self-interested participants. For example, they can be used to elicit a truthful answer to the question "Have you ever used illegal drugs?" It is important to emphasize that peer prediction mechanisms cannot use ground truth for incentive alignment. In the public health example this means that the program has no way of testing whether a participant indeed has or has not ever used illegal drugs. All it can use are the participants' voluntary reports.

The classical *peer prediction method* by Miller et al. [2005] provides an approach to peer prediction. It compares the reported information of a participant with the reported information of another participant, and computes a payment rule that implements truth revelation in a strategic equilibrium. If the participants' prior beliefs are such that every participant always believes that her true answer or experience is also the most likely for others, simply paying participants a reward upon agreement (and nothing otherwise) is sufficient. However, the most interesting peer prediction settings are those where participants may believe that their true answer or experience is not the most likely for others. For example, somebody who has used illegal drugs may still believe that the majority of people have not. What is required for incentive alignment in peer prediction is that there is a positive correlation between an individual's own experience and the experiences of others. For example, a participant may believe that another participant's possibility of having used illegal drugs is 40% if she herself has used drugs and 20% otherwise if she has not.

The major shortcoming of the classical peer prediction method with regard to practical applications is that it relies on too much common knowledge. In particular, the participants' prior beliefs are assumed to be known and the same for all participants. In the public health example, this would mean that every participant using drugs has the same belief that 40% of participants use drugs, and every participant not using drugs has the same belief that 20% do. Moreover, these numbers need to be known by the mechanism in order to compute the payment rule.

In this thesis, I relax these common knowledge assumptions, and design

peer prediction mechanisms that do not require having full knowledge of the participants' prior beliefs. I refer to this property as *robustness* [Bergemann and Morris, 2005].

1.2 Related Areas

It is instructive to differentiate between peer prediction and related research areas:

- Machine Learning approaches.

  Machine learning techniques have been adopted for information aggregation in applications where multiple participants provide noisy reports ("labels") about the same item. Raykar et al. [2010], for example, train a classifier to decide whether a tumor on a medical image is malignant (cancerous) or benign. They present crowd workers on Amazon Mechanical Turk with pictures of tumors and ask them to report the size and shape of the tumor. Based on these reports, they apply expectation maximization to jointly fit the parameters of the classifier, each worker's reliability (or noise level), and the estimated true label. Similarly, Piech et al. [2013] apply machine learning to peer grading in massive open online courses (MOOCs), where students grade each others' assignments.

  Both applications are related to peer prediction because the authors do not assume access to any gold standard or objective ground truth. Moreover, the models used in these papers are more complex than in standard peer prediction mechanisms, in that agents are assumed to differ in reliability, and allow for agent-specific biases. But these approaches do not provide incentives for agents to invest effort or report truthfully. Unifying peer prediction with machine learning is an important direction of future work.

- Classical Mechanism Design and Social Choice.

  Mechanism design is concerned with eliciting private information about preferences [Nisan, 2007]. The outcome of a mechanism is the selected alternative, such as a public choice (e.g., whether to build a bridge) or the allocation of a resource, and it can include payments. Peer prediction is concerned with eliciting more general private information about the agents' environment, such as an agent's experience with a service provider (e.g., the speed of a server or the perceived quality of a hotel). Peer prediction mechanisms always use payments.

  Social choice is concerned with aggregating private information about preferences into a single preference order of a set of alternatives (or to choose a

single alternative). Similar to classical mechanisms of mechanism design, a social choice rule also selects an alternative. However, no payments are used in social choice, and there is less focus on incentives due to the early impossibility results by Arrow [1963], and Gibbard [1973] and Satterthwaite [1975]. In that sense, peer prediction is closer to mechanism design than it is to social choice.

- Prediction Markets and Proper Scoring Rules.

  Prediction markets [e.g. Wolfers and Zitzewitz, 2004; Pennock and Sami, 2007; Chen and Pennock, 2010] and proper scoring rules [e.g. Brier, 1950; Good, 1952] are related to peer prediction in that the focus is to truthfully elicit information from rational, selfish agents using payments. (Prediction markets are also concerned with the aggregation of the elicited information.) The key difference to peer prediction is that these other types of information elicitation mechanisms elicit agents' private probabilistic beliefs about publicly-observable future events. For example, one could use a prediction market or a proper scoring rule to elicit the probabilistic belief that the next president of the United States is a member of the Democratic party. Eventually, it will be publicly known if the event materialized, which is then used to score the agents' probabilistic predictions. That is, ground truth is eventually revealed. This is in contrast to peer prediction, where there is no such publicly-observable event that can be used for scoring but where only the reports of other agents can be used for scoring. Moreover, prediction markets and proper scoring rules are used to elicit probabilistic beliefs, whereas peer prediction is concerned with the elicitation of opinions, ratings, or experiences.

- Principal-Agent and Contract Theory.

  In the economics literature, principal-agent problems and contract theory are areas related to peer prediction in that they also study incentive problems with costly effort and payments. The main difference is that the principal usually observes a noisy signal of an agent's effort. This is in contrast to peer prediction mechanisms, which do not assume any observations on behalf of the mechanism other than through voluntary reports.

Other, more closely related work will be discussed in more detail in the respective chapters.

1.3 Contributions

The contributions in this thesis can be grouped into three parts. In the first part (Chapters 3 and 4), I design the *Robust Bayesian Truth Serum (RBTS)*, which asks agents for two types of reports, the report of the private information it is interested in and a prediction report corresponding to a agent's belief about the private information of other agents. RBTS dispenses with the assumption that the agents' prior beliefs need to be known to compute the payments. It does, however, still rely on all agents having the same prior beliefs. In the second part of the thesis (Chapters 5 and 6), I design *subjective-prior peer prediction mechanisms*, which further reduce the common knowledge assumptions of peer prediction. As in RBTS, subjective-prior mechanisms do not assume knowledge of the agents' prior beliefs. In addition, they allow the agents' prior beliefs to be subjective, i.e. different from one another. In the third part (Chapter 7), I study *effort-incentivizing peer prediction*. In many applications, the information that the designer seeks to elicit first needs to be acquired. When this information acquisition requires costly effort, agents may have an incentive to avoid it and choose to guess instead. I design a payment rule that incentivizes only those agents with good enough information to participate, invest effort, and report their information truthfully, and where all others pass.

Part 1: The Robust Bayesian Truth Serum

The *Bayesian Truth Serum* [Prelec, 2004] is a peer prediction mechanism that does not assume knowledge of the agents' prior beliefs. It asks for two reports: a report about the information itself (the opinion, rating, or experience, which I henceforth refer to as *signal*) and a prediction report corresponding to the agent's belief about the distribution of signals in the population. In the drug example this would mean that each participant, in addition to reporting whether she has used illegal drugs would also have to report her belief that another, randomly-chosen participant is a drug user, e.g. 20%. The mechanism's major drawback is that it is truthful only for a large number of participants, where this number depends on the agents' prior beliefs and is thus unknown to the mechanism.

I design a *Robust Bayesian Truth Serum* (RBTS) that alleviates this problem. As in the original Bayesian Truth Serum, RBTS takes a signal and a prediction report, and does not need to know the participants' prior beliefs. However, RBTS is truthful for three or more participants. For binary signals, I present a version of RBTS that only requires two participants.

RBTS also improves upon BTS in a number of other ways. First, BTS does not satisfy *interim* individual rationality, which means that participants,

after learning their signal, sometimes do not have an incentive to participate in the mechanism because their expected payment is negative. RBTS satisfies the stronger *ex post* individual rationality, meaning that no participant makes a negative payment in any outcome. This is important for many crowdsourcing applications, where it is often infeasible for the mechanism to receive payments from participants. Second, RBTS is well defined for all possible reports, including out-of-equilibrium prediction reports. This is in contrast to BTS, which computes payments of negative infinity when a participant predicts 0% for any signal. Third, RBTS is conceptually simpler than BTS, and the incentive analysis is more straightforward. Moreover, the payments computed by RBTS are bounded for all reports, and this bound can be set to any value chosen by the designer. Bounded ex post payments are important in practice because they provide an upper bound on the designer's willingness to pay. For example, the designer may want to cap payments for a single report at \$0.50.

Part 2: Subjective-Prior Peer Prediction

In the second part of the thesis, I design *subjective-prior peer prediction mechanisms*, which further reduce the common knowledge assumptions of peer prediction. While Bayesian Truth Serum mechanisms do not require knowledge of the agents' prior beliefs, every agent is still assumed to have the same prior beliefs.

Relaxing this assumption of a common prior, I design truthful peer prediction mechanisms where every agent is allowed to have her own *subjective* prior beliefs, i.e. different agents can hold different beliefs despite the same experience. In the aforementioned drug example, this could mean that a participant who uses drugs believes that 40% of participants do, and that another participant who uses drugs believes that 60% do. Obtaining robustness for such differences in beliefs seems important for practical applications.

I first design mechanisms that ask a participant for two reports: one before she observes her signal and one afterwards. The ability to elicit relevant information from a participant both before and after she receives her signal is critical for this first group of subjective-prior mechanisms but seems reasonable in many applications. For example, a travel site could ask a user for her expectation about a hotel at the time of booking and then again after her stay. The *basic subjective-prior peer-prediction mechanism (BSPP)* requires a participant to report two belief reports, one before and one after receiving her signal. BSPP infers the participant's signal from the change in the participant's belief reports. Building on this, I introduce the *shadow subjective-prior peer-prediction mechanism (SSPP)* that, instead of requiring two belief reports, requires only a belief and a signal report.

In moving from a common knowledge model to one with unknown and subjective priors, an important consideration is the amount of additional information that must be elicited from a participant over and above a signal report. Exploring the perceived trade-off between the robustness and reporting costs, I design the *Empirical Shadowing Method*, which allows for subjective priors but requires only a signal report. This combination is compelling because it provides robustness against strategic participants with non-standard (and possibly wrong) beliefs, without requiring truthful participants to deliberate about their beliefs. Moreover, this mechanism dispenses with the requirement that relevant information needs to be elicited both before and after a participant receives her signal.

Part 3: Effort Incentives

Peer prediction mechanisms have traditionally been focused on providing incentives for truthful reporting. However, in many applications, the information that the designer seeks to elicit first needs to be acquired. When this acquisition of information requires costly effort, agents can have an incentive to avoid it and choose to guess instead.

In the classical peer prediction method, the assumption is that mechanism knows the agents' prior beliefs. This allows the method to scale its payments until the expected informational improvement associated with a signal acquisition outweighs the cost for effort. In robust peer prediction, the agents' prior beliefs are no longer assumed to be known, which inhibits the mechanism in computing the required scaling of payments.

I present an approach to address this challenge in a setting that is different from the usual peer prediction setting. It is simpler in that every worker believes that her own experience is also the most likely experience for others but it is more complex in that it incorporates the reliability of an agent. I then explicitly model the decision of an agent as to whether she participates in the mechanism or "passes," in which case she obtains zero utility. In addition, I allow for negative payments. These two changes go hand in hand since every rational agent would choose to participate without negative payments. The main result is that a payment rule can be computed that incentivizes only those agents with good enough information to participate, invest effort, and report their information truthfully, and where all others pass. That is, agents self-select according to quality and those agents with high quality also invest effort. This mechanism is the first peer prediction mechanism that incentivizes fixed-cost effort without the mechanism needing to know the agents' prior beliefs.

Summary of Contributions

To summarize, this thesis makes the following contributions:

- Design of first strictly truthful peer prediction mechanism that does not rely on knowledge of the common belief model to provide strict incentive compatibility for any number of participants $n \geq 3$.

- Design of first Bayesian Truth Serum with bounded ex post payments. Also first numerically stable Bayesian Truth Serum for any inputs, including out-of-equilibrium prediction reports assigning 0% to some signals.

- Natural generalization of classical peer prediction method to robust peer prediction in that participants are presented with a menu of signals that is translated into a belief report, which is then scored using a proper scoring rule.

- First analysis of how different conditions on the signal prior and signal posterior required for strict truthfulness by different peer prediction mechanisms relate to each other.

- First definition of a solution concept for subjective priors. This *ex post subjective equilibrium* is more general than Bayes-Nash equilibrium but less general than dominant-strategy implementation.

- Design of first strictly truthful peer prediction mechanism for a model without a common prior.

- First analysis of peer prediction mechanism that requires sequential information elicitation.

- First peer prediction mechanism that combines minimal reporting, i.e. participants only need to report their signals, with subjective priors; laying out theory for learning the signal prior.

- Development of general method that incentivizes fixed-cost effort when mechanism is not assumed to know participants' belief models. When participants differ in reliability/quality, method provides free screening through self-selection of participants.

- Discovery and exploration of trade-off between robustness of incentive properties, reporting requirements, and number of participants.

1.4 Bibliographic Note

This thesis is based on the following publications. I am the first author on all of them.

1. Witkowski, J. and Parkes, D. (2012a). A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, pages 1492–1498.

2. Witkowski, J. and Parkes, D. (2012b). Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*, pages 964–981.

3. Witkowski, J. and Parkes, D. (2013). Learning the Prior in Minimal Peer Prediction. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC'13)*.

4. Witkowski, J., Bachrach, Y., Key, P., and Parkes, D. (2013). Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*, pages 190–197.

The following publications were published during my doctoral studies but are not part of the thesis.

1. Witkowski, J. (2009). Eliciting Honest Reputation Feedback in a Markov Setting. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09)*, pages 330–335.

2. Witkowski, J. (2010). Truthful Feedback for Sanctioning Reputation Mechanisms. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI'10)*, pages 658–665.

3. Witkowski, J., Seuken, S., and Parkes, D. (2011). Incentive-Compatible Escrow Mechanisms. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, pages 751–757.

4. Witkowski, J. (2011). Incentive-Compatible Trust Mechanisms (Extended Abstract). In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, pages 1865–1866.

5. Witkowski, J. (2011). Trust Mechanisms for Online Systems (Extended Abstract). In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, pages 2866–2867.

1.5 Outline

The remainder of the thesis is organized as follows. In Chapter 2, I introduce the classical peer prediction method and its common knowledge belief model. From Chapter 3 to Chapter 5, I successively relax these knowledge assumptions. In Chapter 3, I first introduce the Shadowing Method, which requires less common knowledge than the classical peer prediction method. The Shadowing Method also serves as a building block for mechanisms in later chapters, including the Robust Bayesian Truth Serum (RBTS), which I introduce in Chapter 4. In Chapter 5, I then introduce truthful mechanisms that allow for subjective prior beliefs, which are also unknown to the mechanism. In Chapter 6, I present the Empirical Shadowing Method, which, as the mechanisms from Chapter 5, allows for subjective prior beliefs, but which only requires signal reports. In Chapter 7, I analyze a setting that is different from the setting studied in Chapters 2 to 6, in that the participants' signals are of different quality and require the investment of effort associated with a known, fixed cost. I conclude with a brief discussion of the lessons learned and the most important directions of future work in Chapter 8.

Chapter 2

Classical Peer Prediction

In this chapter, I introduce the *peer prediction method* by Miller et al. [2005], which is the classical mechanism to incentivize truthful reporting in peer prediction. The major shortcoming of classical peer prediction is that it critically relies on common knowledge assumptions that are prohibitive in practice. In the drug example introduced in Chapter 1, this mechanism would require that all participants who use drugs have the same belief about a randomly-chosen participant being a drug user (e.g. 40%) and all participants not using drugs have the same belief about a randomly-chosen participant being a drug user (e.g. 20%). Moreover, to compute the payment rule, the classical peer prediction method needs to assume that the mechanism knows these beliefs as well.

The remainder of this chapter is organized as follows: in Section 2.1, I describe the basic peer prediction model and I give a numerical example in Section 2.2. In Section 2.3, I introduce the required game-theoretic concepts, followed by an analysis of the simple output agreement mechanism in Section 2.4. After an introduction to proper scoring rules, I present Miller et al.'s classical peer prediction mechanism, the *peer prediction method*, in Section 2.5. Section 2.6 provides a discussion of mechanisms in the classical peer prediction model, and Section 2.7 concludes the chapter with a brief discussion of the shortcomings and challenges of classical peer prediction.

2.1 Model

There is a group of $n \geq 2$ rational, risk-neutral and self-interested agents. A *world state* is determined by random variable T that can adopt values in the set $\{1, \ldots, l\}$. When interacting with the world, each agent i observes a signal S_i, which is a random variable with values $\{1, \ldots, m\}$. The signal represents an agent's experience or opinion. The objective in peer prediction is to elicit

an agent's signal in an incentive compatible way, i.e. to compute payments such that agents maximize their expected payment by reporting their signal to the mechanism (center) truthfully.

In the classic set-up, all agents share a common belief model in regard to the state of the world and the distribution of signals conditioned on world states. That is, every agent i has the same belief $\Pr(T = t)$ about the world state before observing a signal, and the same conditional belief $\Pr(S = s \mid T = t)$ for how signals are generated for each possible state. (I will use random variable S as a generic signal and s as an instantiation of S.) I will assume that $\Pr(T = t) > 0$ for all $t \in \{1, \ldots, l\}$ as any state with probability zero can be eliminated without changing the model's behavior. It is crucial that this belief model is the same for all agents and, moreover, that it is known by the mechanism.

When an agent observes a signal, she updates her state and signal beliefs according to the belief model. We adopt shorthand

$$p(s_j|s_i) = \Pr(S_j = s_j \mid S_i = s_i) \tag{2.1}$$

for agent i's signal posterior belief that a second agent j (henceforth agent i's *peer* agent) receives signal s_j given agent i's signal s_i.

The signal posterior can be calculated as

$$p(s_j|s_i) = \Pr(S_j = s_j \mid S_i = s_i) = \sum_{t=1}^{l} \Pr(S_j = s_j \mid T = t) \Pr(T = t \mid S_i = s_i). \tag{2.2}$$

Applying Bayes' rule to the second part of the summation in Equation 2.2 yields:

$$\Pr(T = t \mid S_i = s_i) = \frac{\Pr(S_i = s_i \mid T = t) \Pr(T = t)}{\Pr(S_i = s_i)}. \tag{2.3}$$

The denominator in Equation 2.3 is the signal prior belief and can be computed as

$$\Pr(S_i = s_i) = \sum_{t=1}^{l} \Pr(S_i = s_i \mid T = t) \Pr(T = t). \tag{2.4}$$

Similar to the signal posteriors, we denote the signal prior for signal s_i by

$$p(s_i) = \Pr(S_i = s_i). \tag{2.5}$$

Note that in this classic set-up, agents are differentiated only by the signal they receive. In particular, it holds that $\Pr(S_j = s'|S_i = s) = \Pr(S_i = s'|S_j = s)$ for all $s, s' \in \{1, \ldots, m\}$.

To describe belief models, I use vector and matrix representations that allow a compact representation. In a vector, row number k corresponds to the belief

for state k or signal k. Matrices represent conditional probabilities and are read as row given column from left to right and top to bottom. That is, a belief model is given by:

$$\Pr(T) = \begin{pmatrix} \Pr(T=1) \\ \cdots \\ \Pr(T=l) \end{pmatrix}$$

$$\Pr(S|T) = \begin{pmatrix} \Pr(S=1|T=1) & \ldots & \Pr(S=1|T=l) \\ \ldots & \ddots & \vdots \\ \Pr(S=m|T=1) & \ldots & \Pr(S=m|T=l) \end{pmatrix}.$$

This results in signal prior

$$p(\cdot) = \Pr(S|T) \times \Pr(T) = \begin{pmatrix} \sum_{t=1}^{l} \Pr(S=1|T=t) \cdot \Pr(T=t) \\ \cdots \\ \sum_{t=1}^{l} \Pr(S=m|T=t) \cdot \Pr(T=t) \end{pmatrix}$$

and, using Equations 2.2 and 2.3, in signal posterior matrix

$$p(\cdot|\cdot) = \begin{pmatrix} p(1|1) & \ldots & p(1|m) \\ \ldots & \ddots & \vdots \\ p(m|1) & \ldots & p(m|m) \end{pmatrix}.$$

Furthermore, for the signal prior and the signal posteriors it holds that

$$p(s) = \sum_{k=1}^{m} p(s|k) \cdot p(k) \tag{2.6}$$

and thus

$$\begin{pmatrix} p(1|1) & \ldots & p(1|m) \\ \ldots & \ddots & \vdots \\ p(m|1) & \ldots & p(m|m) \end{pmatrix} \times \begin{pmatrix} p(1) \\ \cdots \\ p(m) \end{pmatrix} = \begin{pmatrix} p(1) \\ \cdots \\ p(m) \end{pmatrix}.$$

Another way of saying this is that the signal prior is an eigenvector of the signal posterior matrix with eigenvalue 1.

A crucial assumption for the existence of strictly incentive compatible peer prediction mechanisms is *stochastic relevance* [Johnson et al., 1990].

Definition 1. Random variable S_i is *stochastically relevant* for random variable S_j if and only if the distribution of S_j conditional on S_i is different for all possible realizations of S_i.

That is, stochastic relevance holds if and only if $p(\cdot|s) \neq p(\cdot|s')$ for all $s' \neq s$, i.e. if all columns in the signal posterior matrix are different. Miller et al. [2005] show that small belief perturbations make stochastically irrelevant belief combinations stochastically relevant. While it is standard in the peer prediction literature to assume that different world states t induce different signal distributions, i.e. $\Pr(S = s \mid T = t') \neq \Pr(S = s \mid T = t)$ for all $t', t \in \{1, \ldots, l\}$ with $t' \neq t$, this is not sufficient for stochastic relevance.

Going forward, it is assumed that $p(s) > 0$ for all $s \in \{1, \ldots, m\}$. This is without loss of generality because, if there is a signal s for which $p(s) = 0$, we can just remove the signal from the model.

2.2 Example

In this section, I present an example of a peer prediction setting that I will use throughout the thesis. It is motivated by a challenge in online advertising. Display ads (also called "banner ads") are displayed next to a website's primary content. They are often sold through an intermediary (a so-called *ad exchange*) between the advertiser and the content provider [e.g. Goldstein et al., 2012]. For reasons of brand reputation, advertisers care about the website their ads are placed on. Premium brands typically do not want to advertise on websites with offensive content, such as nudity or violence. While ad exchanges allow advertisers to specify the types of websites they want, content providers displaying nudity or violence have no incentive to reveal this to the ad exchange because this would reduce competition for their ad slots, and thus also reduce revenue. Moreover, these content providers can also hide the offensive nature of their content from algorithms by avoiding certain keywords.

Instead of exclusively relying on self reports or keyword-based algorithms, ad exchanges can turn to crowdsourcing platforms, such as Amazon Mechanical Turk, and ask crowd workers to report whether a given website contains offensive content.[1] Reporting accurate information requires workers to invest effort because they need to go through the website and compare its content with the detailed definition of what is considered offensive. This is time-consuming. When workers are paid the same amount for all reports, however, their incentive is to go through them as quickly as possible without investing due effort.

[1] For example, see the methodology used by Integral Ad Science at `http://integralads.com/our-technology/rating-methodology` and a blog post on the same topic at `http://www.behind-the-enemy-lines.com/2013/06/facebook-implements-brand-safety-doing.html`.

Peer prediction mechanisms can address this challenge of incentivizing workers to invest effort by providing payments that depend not only on the report of the worker herself but also on the report of another worker investigating the same website.

Example 1. *An ad exchange presents two workers with the same website and asks each of them whether the website does contain violence. The website is in one of only two possible states, $T = 1$ ("no violence") and $T = 2$ ("violence"). Similarly, the signals $S = 1$ and $S = 2$ correspond to "no violence observed" and "violence observed," respectively. Furthermore, let the common belief model be given by*

$$\begin{aligned} \Pr(T = 2) &= 0.3 \\ \Pr(S = 2 \mid T = 2) &= 0.6 \\ \Pr(S = 2 \mid T = 1) &= 0.1 \end{aligned}$$

The values for $\Pr(T = 1) = 1 - \Pr(T = 2)$, $\Pr(S = 1 \mid T = 2) = 1 - \Pr(S = 2 \mid T = 2)$, and $\Pr(S = 1 \mid T = 1) = 1 - \Pr(S = 2 \mid T = 1)$ follow directly because there are only two states and two signals.

Given this belief model, the signal prior that an agent will observe violence (signal 2) is

$$\Pr(S = 2) = \Pr(S = 2 \mid T = 2)\Pr(T = 2) + \Pr(S = 2 \mid T = 1)\Pr(T = 1) = 0.25.$$

By viewing the website, the worker learns something about its state. For example, following signal $S_i = 2$ ("violence observed"), worker i updates her belief that the website does in fact contain violence to

$$\Pr(T = 2 | S_i = 2) = \frac{\Pr(S_i = 2 | T = 2)\Pr(T = 2)}{\Pr(S_i = 2)} = 0.72$$

The analogous update following $S_i = 1$ ("no violence observed") is $\Pr(T = 2 | S_i = 1) = 0.16$. Even after observing a signal, the belief that the website contains violence is neither 0% nor 100% because of noise in the worker's signal and the inevitably incomplete definition of what constitutes violence (a large knife, for example, may or may not be interpreted as a weapon).

Given this updated belief regarding the state of the website, the worker revises her belief that her peer worker j, who is looking at the same website, observed violence. For example, given $S_i = 2$ ("violence observed"), worker i's signal

posterior that worker j also observed violence is:

$$\begin{aligned} p(2|2) = &\quad \Pr(S_j = 2 \mid T = 1) \Pr(T = 1 \mid S_i = 2) \\ &+ \Pr(S_j = 2 \mid T = 2) \Pr(T = 2 \mid S_i = 2) \\ = &\quad 0.46 \end{aligned}$$

The analogous update following $S_i = 1$ is $p(2|1) = 0.18$.

Observe that $p(2|2) < p(1|2)$, so that, even after observing violence, worker i still believes that it is more likely that her peer worker j did not observe violence. This can, for example, be the case when worker i observes an obstructed knife which, she believes, worker j will not see. In fact, these difficult cases are also the most interesting instances of human computation tasks, because easier cases can likely be solved using computer vision algorithms.

2.3 Game-Theoretic Concepts

Before describing peer prediction mechanisms, I first introduce some basic game-theoretic concepts. I will write index $-i$ to denote all agents except for agent i. For example, S_{-i} denotes the vector of the $n-1$ signals received by all agents but agent i. Note that, as before, S and s are not vectors but a generic signal and its instantiation, respectively.

Definition 2. A *pure strategy* $\sigma_i : \{1, \ldots, m\} \to \{1, \ldots, m\}$ for agent i describes her signal report for every possible signal observation.

Peer prediction mechanisms pay participants a score, and it is assumed that agent i's utility u_i is linear in this score.

Definition 3. The *utility* $u_i\big(\sigma_i(s_i), \sigma_{-i}(s_{-i})\big)$ of agent i depends on the vector of reported signals.

The notation of u_i is overloaded in cases where the score of agent i depends only on the reports of a subset of agents. In that case, I will give only the strategies of the agents in this subset. For example, when u_i depends only on the reports of agents i and j, I will write $u_i\big(\sigma_i(s_i), \sigma_j(s_j)\big)$.

The equilibrium concept adopted for the analysis of peer prediction mechanisms is that of a Bayes-Nash equilibrium:[2]

[2]One could instead refer to this as a correlated equilibrium, since each agent observes a correlated signal from nature, and because the agents' utilities only depend on the reports of agents. I opt for the more standard Bayes-Nash equilibrium terminology.

Definition 4. Strategy profile $(\sigma_1^*, \ldots, \sigma_n^*)$ is a *Bayes-Nash equilibrium (BNE)* of a peer prediction mechanism with n agents if

$$\mathbf{E}_{S_{-i}}\Big[u_i\big(\sigma_i^*(s_i), \sigma_{-i}^*(S_{-i})\big) \mid S_i = s_i\Big] \geq \mathbf{E}_{S_{-i}}\Big[u_i\big(\sigma_i(s_i), \sigma_{-i}^*(S_{-i})\big) \mid S_i = s_i\Big]$$

for all $i \in \{1, \ldots, n\}$, all $s_i \in \{1, \ldots, m\}$, and all $\sigma_i \neq \sigma_i^*$. It is a strict BNE if the inequality is strict.

Each agent maximizes her expected score by following strategy σ_i^* given her own signal and given that the other agents play according to strategies σ_{-i}^*.

Definition 5. In the *truthful strategy*, strategy $\sigma_i(s_i) = s_i$ for each $s_i \in \{1, \ldots, m\}$.

Definition 6. A mechanism is *Bayes-Nash incentive compatible (BNIC)* if it is a BNE for all agents $i \in \{1, \ldots, n\}$ to play the truthful strategy. A mechanism is *strictly* BNIC if this is a strict BNE.

We will sometimes refer to (strictly) BNIC mechanisms as (strictly) truthful mechanisms.

So far, we have assumed that agents already observed their signals when participating in the mechanism. However, peer prediction mechanisms are especially useful for incentivizing effort, i.e. the costly acquisition of signals. For an example of such a setting, see Section 2.2. Let each agent i incur cost C for obtaining her signal S_i. If no effort is invested, the agent's belief about another agent j's signal is her signal prior.

Definition 7. A strictly BNIC peer prediction mechanism *implements effort cost C* if and only if for every agent $i \in \{1, \ldots, n\}$, the expected utility—given that all other agents invest effort and report truthfully—is higher when agent i invests effort than when she invests no effort, i.e. it holds that

$$\mathbf{E}_{S_i, S_{-i}}\Big[u_i(S_i, S_{-i})\Big] - C > \max_{\hat{s_i} \in \{1, \ldots, m\}} \mathbf{E}_{S_{-i}}\Big[u_i(\hat{s_i}, S_{-i})\Big]$$

where $\hat{s_i}$ is agent i's signal report that maximizes her expected utility according to the signal prior.

Theorem 2.1. *If a mechanism $\mathcal{M}$ is strictly BNIC, then $\mathcal{M}$ also incentivizes effort cost C for some $C > 0$.*

Proof. All that needs to be shown is that the strict inequality in Definition 7 holds for $C = 0$ because if strictness holds, there is always some $\varepsilon > 0$ that can be subtracted from the left-hand side, so that the inequality still holds. Intuitively, it needs to be shown that an agent faced with a strictly truthful

mechanism is strictly better off obtaining information when it comes for free. Let

$$\hat{s} = \arg\max_{\hat{s_i} \in \{1,\ldots,m\}} \mathbf{E}_{S_{-i}}\Big[u_i(\hat{s_i}, S_{-i})\Big]$$

be the signal report maximizing agent i's expected utility without investing effort and using only the signal prior. One then obtains

$$\begin{aligned}
\mathbf{E}_{S_{-i}}\Big[u_i(\hat{s}, S_{-i})\Big] &= \sum_{s_{-i}} \Pr(S_{-i} = s_{-i}) \cdot u_i(\hat{s}, s_{-i}) \\
&= \sum_{s_{-i}} \sum_{s_i} \Pr(S_{-i} = s_{-i} \wedge S_i = s_i) \cdot u_i(\hat{s}, s_{-i}) \\
&= \sum_{s_i} \Pr(S_i = s_i) \cdot \sum_{s_{-i}} \Pr(S_{-i} = s_{-i} | S_i = s_i) \cdot u_i(\hat{s}, s_{-i}) \\
&< \sum_{s_i} \Pr(S_i = s_i) \cdot \sum_{s_{-i}} \Pr(S_{-i} = s_{-i} | S_i = s_i) \cdot u_i(s_i, s_{-i}) \\
&= \mathbf{E}_{S_i}\left[\mathbf{E}_{S_{-i}}\Big[u_i(s_i, S_{-i}) \mid S_i = s_i\Big]\right] \\
&= \mathbf{E}_{S_i, S_{-i}}\Big[u_i(S_i, S_{-i})\Big]
\end{aligned}$$

where the strict inequality follows from the strict truthfulness of $\mathcal{M}$. □

Theorem 2.1 is important because it allows us to focus on the design of strictly truthful mechanisms. Observe that strict incentives for truthfulness are crucial for this result. This is why weakly truthful mechanisms, such as paying every agent a fixed amount that is independent of the agent's report, are not sufficient.

2.4 Output Agreement

The simplest mechanism in the peer prediction space is simple output agreement, where two agents report their signals and are paid a fixed amount if and only if their reports agree. I introduce and analyze this mechanism in this section.

2.4.1 Mechanism

Simple output agreement is defined as:

1. Each agent i is asked for her signal report $x_i \in \{1, \ldots, m\}$.

2. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$ and pay agent i:

$$u_i(x_i, x_j) = \begin{cases} \tau & \text{if } \; x_i = x_j \\ 0 & \text{if } \; x_i \neq x_j \end{cases}$$

where $\tau > 0$ and x_j is the signal report by *peer* agent j.

For example, given two agents, an instance of a simple output agreement mechanism is to pay \$1 to each agent when both agents report the same signal, and \$0 otherwise.

2.4.2 Incentive Analysis

Theorem 2.2. *Simple output agreement is strictly BNIC if and only if* $p(s|s) > p(s'|s)$ *for all signals* $s, s' \in \{1, \ldots, m\}$ *with* $s' \neq s$.

Proof. Since agent i's utility only depends on her own report and the report of her peer agent j, it is sufficient to consider only these two agents. For all signals $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$, we have:

$$\begin{aligned} & \mathbf{E}_{S_j}\Big[u_i(s, S_j) \mid S_i = s\Big] > \mathbf{E}_{S_j}\Big[u_i(s', S_j) \mid S_i = s\Big] \\ \Leftrightarrow \quad & p(s|s) \cdot \tau > p(s'|s) \cdot \tau \Leftrightarrow p(s|s) > p(s'|s) \end{aligned}$$ □

This means that, since it only matters that the signal reports agree (but not on which signal), agent i's unique best response is to report the signal that is most likely to be agent j's signal. In particular, simple output agreement is not truthful in settings where people can hold minority opinions or experiences. For example, when scored using simple output agreement and asked whether Chicago is the capital of Illinois, a rational agent knowing that Chicago is not the capital will still report that it is if she believes that the majority of other agents believes that Chicago is the capital.

Note that multiple equilibria are unavoidable in strictly truthful peer prediction mechanisms [Jurca and Faltings, 2005, 2009; Waggoner and Chen, 2013]. For example, in simple output agreement, all agents reporting signal 1 is also an equilibrium.

Example 2. *Consider again the belief model from Example 1 from page 15, where* $p(2|2) = 0.46 < 0.54 = p(1|2)$. *Suppose simple output agreement was applied, and agent* i *observed signal* $S_i = 2$. *Assuming that agent* j *truthfully reported her signal, agent* i*, in trying to match agent* j*'s report, would not be truthful, because her expected utility for lying and reporting* 1 *would be* 0.54τ*, whereas her expected utility for reporting* 2 *would only be* 0.46τ. *Furthermore,*

since $p(1|1) = 0.82 > 0.18 = p(2|1)$ *in that example, agent* i *would maximize her expected utility by always reporting signal* $x_i = 1$*, independent of her true signal.*

2.5 The Peer Prediction Method

In this section, I introduce the classical peer prediction method developed by Miller, Resnick, and Zeckhauser [2005]. The advantage of this mechanism when compared to simple output agreement is that it also works for belief models where agents believe to have a minority opinion. Since the classical peer prediction method relies on proper scoring rules, I introduce them first.

2.5.1 Proper Scoring Rules

Proper scoring rules can be used to incentivize a rational agent to truthfully report her private, probabilistic belief about the likelihood of a future event.

Let $\mathcal{D}$ denote the set of probability distributions over a set of outcomes, where the outcomes are clear from context. Let $b \in \mathcal{D}$ be the agents probabilistic belief about each outcome. The timing is then as follows: first, the agent is asked for her belief report $y \in \mathcal{D}$. Second, an event $\omega \in \Omega$ materializes (observed by the mechanism) and, third, the agent receives payment $R(y, \omega)$.

Definition 8 (Scoring Rule)**.** Given possible outcomes $\Omega = \{1, \ldots, m\}$, and a report $y \in \mathcal{D}$ in regard to the probability distribution over Ω, a scoring rule $R(y, \omega) \in \mathbb{R} \cup \pm\infty$ assigns a score based on report y and the outcome ω that occurs.

Definition 9 (Strictly Proper Scoring Rule)**.** A scoring rule is *proper* if an agent maximizes her expected score by truthfully reporting her belief $b \in \mathcal{D}$, and is *strictly proper* if the truthful report is the only report that maximizes the agent's expected score.

An example of a strictly proper scoring rule is the logarithmic scoring rule $R_{\log}$

$$R_{\log}(y, \omega) = \ln(y(\omega)). \tag{2.7}$$

for natural logarithm ln.

Proposition 2.3. *[Good, 1952] The logarithmic scoring rule* $R_{\log}$ *is strictly proper.*

A positive-affine transformation of a proper scoring rule is still proper.

Another example of a proper scoring rule is the quadratic scoring rule that I discuss in more detail in Section 3.4. For an in-depth treatment of proper scoring rules in general, I refer to the article by Gneiting and Raftery [2007].

2.5.2 Mechanism

The *classical peer prediction method* is defined as:

1. Each agent i is asked for her signal report $x_i \in \{1, \ldots, m\}$.
2. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$, and use knowledge of $p(\cdot|\cdot)$ to pay agent i

$$u_i(x_i, x_j) = R\big(p(s_j|x_i), x_j\big),$$

where R is a proper scoring rule and x_j the signal report by agent j.

Because the belief model is known by the mechanism, $p(s_j|x_i)$ can be computed for any x_i. Knowing $p(s_j|x_i)$, the score $R\big(p(s_j|x_i), x_j\big)$ can be computed by the mechanism for all $x_i, x_j \in \{1, \ldots, m\}$ as well.

2.5.3 Incentive Analysis

Theorem 2.4. *[Miller et al., 2005] The classical peer prediction method is strictly BNIC for any strictly proper scoring rule R, given that the common belief model satisfies stochastic relevance.*

Proof. Since agent i's utility only depends on her own report and the report of her peer agent j, it is sufficient to consider only these two agents. For all signals $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$, we have:

$$\begin{aligned}
& \mathbf{E}_{S_j}\Big[u_i(s, S_j) \;\Big|\; S_i = s\Big] > \mathbf{E}_{S_j}\Big[u_i(s', S_j) \;\Big|\; S_i = s\Big] \\
\Leftrightarrow \quad & \sum_{s_j=1}^{m} p(s_j|s) \cdot R\big(p(s_j|s), s_j\big) > \sum_{s_j=1}^{m} p(s_j|s) \cdot R\big(p(s_j|s'), s_j\big) \\
\Leftrightarrow \quad & p(\cdot|s) \neq p(\cdot|s'),
\end{aligned}$$

which is stochastic relevance. Note that the left hand side of the second term is precisely the expected utility of agent i if she reported her true probabilistic belief $p(\cdot|s)$ about S_j to a strictly proper scoring rule R. The right hand side is the expected utility of agent i if she reported belief $p(\cdot|s')$ to a strictly proper scoring rule. If these two beliefs are different, then it follows directly from strict properness of R that the left hand side is larger. □

The intuition is that since the classical peer prediction method knows the belief model of the agents, it also knows the m possible beliefs an agent may have following each of the possible signal observations. The method utilizes this by asking each agent to only report her signal, which the mechanism first transforms into the correct belief report, and which is then applied to a strictly proper scoring rule.

Definition 10. A peer prediction mechanism satisfies *ex post individual rationality (ex post IR)* if all payments are non-negative for all reports.

Proposition 2.5. *Miller et al. [2005] The classical peer prediction method can be made ex post IR for any proper scoring rule R given fully mixed signal posteriors, i.e. $p(s|s') > 0$ for all $s, s' \in \{1, \dots, m\}$.*

In the classical peer prediction method, there is a finite set of signals that can be reported and thus a finite set of possible signal posteriors. The mechanism can compute these signal posteriors (because the belief model is known) and thus the scores that can arise for a given scoring rule. One can thus add a constant to every payment, such that the minimum payment plus this constant is non-negative for any possible combination of signal reports. If not all signal posteriors are fully mixed, a proper scoring rule that has bounded ex post score for belief reports of 0%, such as the quadratic scoring rule that I introduce in Section 3.4, needs to be used.

Example 3. *Consider again the numbers from Example 1 on p. 15 with $m = 2$ possible signals, where the signal posteriors are given by $p(2|2) = 0.46$ and $p(2|1) = 0.18$. Assume that agent 1 observes signal $S_i = 1$, so that her belief about the signal observed by agent 2 is her signal posterior*

$$\begin{pmatrix} p(1|1) \\ p(2|1) \end{pmatrix} = \begin{pmatrix} 0.82 \\ 0.18 \end{pmatrix}.$$

Using the logarithmic rule $R_{\log}$ and assuming agent 2 reports truthfully, agent 1's expected utility reporting truthfully as well is

$$\begin{aligned} \mathbf{E}_{S_2}\Big[u_1\big(x_1 = 1, S_2\big)\Big) \;\Big|\; S_i = 1\Big] &= p(1|1) \cdot u_1(1,1) + p(2|1) \cdot u_1(1,2) \\ &= 0.82 \cdot \ln(0.82) + 0.18 \cdot \ln(0.18) \\ &= -0.47 \end{aligned}$$

If agent 1 *misreports her true signal, i.e.* $x_1 = 2$*, her expected utility is*

$$\begin{aligned}\mathbf{E}_{S_2}\Big[u_1\big(x_1 = 2, S_2\big) \;\Big|\; S_i = 1\Big] =& p(1|1) \cdot u_1(2,1) + p(2|1) \cdot u_1(2,2) \\ =& 0.82 \cdot \ln(0.54) + 0.18 \cdot \ln(0.46) \\ =& -0.65.\end{aligned}$$

That is, given agent 2 *reports truthfully and agent* 1 *observes* $S_1 = 1$*, agent* 1*'s unique best response is to report truthfully as well. The situation is analogous following* $S_1 = 2$*, and since agent* 2 *has the same belief model as agent* 1*, both agents reporting truthfully is a Bayes-Nash equilibrium.*

To ensure ex post individual rationality, the mechanism can use a scaled version of the logarithmic rule $R'_{\log}$*, with an added constant corresponding to the absolute value of the lowest possible negative payment, such that*

$$\begin{aligned}R'_{\log} =& R_{\log} + \big|\min\big(\ln(0.46), \ln(0.54), \ln(0.18), \ln(0.82)\big)\big| \\ =& R_{\log} + |\ln(0.18)| = R_{\log} + 1.71.\end{aligned}$$

2.6 Extensions

Several extensions to the classical peer prediction method have been proposed. Here, I only discuss the extensions using the classical peer prediction method's common knowledge model, and refer to Chapters 3 to 7 and their related work sections for discussions of extensions that relax the common knowledge assumptions.

Jurca and Faltings [2006] formulate the peer prediction method as the solution to a linear program with the objective of minimizing the method's expected payment subject to strict truthfulness. Jurca and Faltings also show how to avoid collusive equilibria [2009]. In my own work, I extend the method to settings where ground truth changes over time, which is common in computational settings, such as users reporting on the speed of server clusters [Witkowski, 2009]. In another line of work, I develop a peer prediction mechanism for buyer feedback on eBay-like online auction sites [Witkowski, 2010] with the interesting aspect that the seller is a strategic player, so that the experience of a buyer is no longer purely stochastic but strategically chosen by the seller.

2.7 Conclusion

In this chapter, I have introduced the classical peer prediction method. Its strength, when compared to other peer prediction mechanisms such as simple output agreement, is that it barely restricts the belief model for which it can

provide strict truthfulness. All that is needed is stochastic relevance, i.e. that the signal posteriors following different signals are different.

The classical peer prediction method has two major shortcomings. First, it critically relies on the assumption of all participants sharing the same belief model, and second, it also assumes this belief model to be known to the mechanism. These assumptions are not satisfied in practice. In the remainder of this thesis, the focus will thus be on peer prediction mechanisms that relax these assumptions.

Note that there are other challenges in peer prediction. First, in addition to the truthful equilibrium, every truthful peer prediction mechanism also has other, non-truthful equilibria [Jurca and Faltings, 2005, 2009; Waggoner and Chen, 2013]. Second, the classical peer prediction method relies on payments, which is not always feasible in online environments. Third, the assumption that the signal observed by one participant is indistinguishable from the signal observed by another participant seems most natural in settings where experiences are in some sense objective. Imagine, for example, a travel website that elicits from users whether their flights were delayed. In this situation, the assumption that users' experiences are drawn from the same distribution seems appropriate as long as it is properly defined what constitutes a delay. In other settings, however, signals are likely to be influenced by user characteristics, such as differences in taste. As an example, consider a reputation system eliciting user feedback about the "quality" of books. Here it seems likely that different readers have different tastes. As the authors of the classical peer prediction method point out, the method can in principle incorporate such user differences through their explicit modeling. The problem with this approach is that it further adds onto the common knowledge that needs to be assumed since this would require different belief models for different types of users, and the assumption that the mechanism knows these, their distribution, and each user's type. This seems unrealistic in practice. Since these challenges not only affect the classical peer prediction method but peer prediction in general, I refer to Chapter 8 for a more detailed discussion.

Chapter 3

The Shadowing Method

Instead of assuming that the entire belief model is common knowledge as in the classical peer prediction method, the Shadowing Method only needs to know the signal prior. To motivate knowledge of the signal prior but not the full model, consider eliciting reports on how noisy a restaurant is, where the platform knows the fraction of times similar restaurants tend to be noisy, or eliciting reports on whether or not a website contains offensive content where the platform knows the fraction of times that similar websites tend to have offensive content. Another setting is where the prediction from a trained classifier has low confidence, e.g., perhaps the prediction is that the site contains offensive content with probability 0.6. This can form the signal prior for the Shadowing Method.

Moreover, the Shadowing Method proves useful as a building block for peer prediction mechanisms that do not assume knowledge about either the signal prior or the signal posteriors. The question when using the Shadowing Method as a building block for these knowledge-free mechanisms is where the signal prior comes from, and the mechanisms of Chapters 4 to 6 each provide a different answer to this question.

Before delving into the Shadowing Method, Section 3.3 presents the 1/prior (read: "one over prior") mechanism [Jurca and Faltings, 2008, 2011] which works under the same knowledge assumptions but requires different conditions to hold for the belief model in order to be truthful. From a technical point of view, the 1/prior mechanism can be seen as a generalization of simple output agreement (Section 2.4), in that the agents are only paid if their signal reports agree. The 1/prior mechanism is more general than simple output agreement because the amounts paid for agreement depend on which signal the agents agree upon.

The Shadowing Method (Sections 3.5 and 3.6), on the other hand, is in the spirit of the classical peer prediction method (Section 2.5) in that both methods follow the same overall procedure of taking a signal report, transforming this

report into a belief report, and then scoring this transformed belief report using a scoring rule. However, since the Shadowing Method only knows the signal prior but not the signal posteriors following each possible signal observation, the technicalities of the Shadowing Method are more intricate than those of the classical peer prediction method.

The Shadowing Method solves two interconnected challenges. It first needs to compute a belief that is in some sense "similar" to an agent's true signal posterior belief, and it then needs to be careful about the scoring of this belief. As we will see in Section 3.5, the Shadowing Method perturbs the signal prior using the agent's signal report in a particular way, and then uses the quadratic scoring rule to score this report. Interestingly, because the perturbed signal prior is not guaranteed to perfectly match the true signal posterior of an agent, the method relies on a scoring rule that is more than just strictly proper, which is in contrast to the classical peer prediction method which works with any strictly proper scoring rule. The reason is that strict properness does not specify an agent's optimal report when the possible reports are only a subset of all distributions and when the agent's true belief is not in that subset. In its current form, the Shadowing Method thus critically relies on the quadratic scoring rule, which has the property that the "closer" a belief report is to the true belief, the higher the agent's expected score.

The remainder of this chapter is organized as follows. In Section 3.1, I briefly discuss related work. In Section 3.2, I explain the difference of the model used in this chapter as compared to the standard model of Chapter 2. In Section 3.3, I then introduce and analyze the 1/prior mechanism due to Jurca and Faltings [2008, 2011]. After introducing the quadratic scoring rule and its properties in Section 3.4, I then present the Shadowing Method for two signals in Section 3.5. In Section 3.6, I then show two different ways of generalizing this binary Shadowing Method to three or more signals, i.e. $m \geq 3$. I compare the 1/prior mechanism with the Shadowing Method in Section 3.7 and conclude the chapter with Section 3.8.

3.1 Related Work

In addition to the related work on peer prediction mechanisms introduced in Chapters 2 to 5, there is additional related work particularly relevant to the mechanisms presented in this chapter.

In Section 3.6.2, I present a Multi-Signal Shadowing Method that is based on a reduction from the Binary Shadowing Method. The idea is conceptually similar to the work by Matheson and Winkler [1976]. They describe a general method for the design of proper scoring rules for continuous probability dis-

tributions that first partitions the real line into two intervals, and then uses a binary proper scoring rule to score the prediction for the real space by mapping it into this binary space. Their work differs from the multi-signal to binary-signal reduction presented here in two aspects. First, instead of belief reports, the Shadowing Method takes signal reports. Second, the random partitioning used for the Binary Shadowing Method reduction takes a particular form, in that each partition consists of a singleton in one group and all other signals in the other group.

3.2 Model

The model of this chapter is the variation of the standard model (Section 2.1), where the belief model is common knowledge amongst all agents, but the mechanism only needs to know the signal prior.

3.3 1/prior Mechanism

The 1/prior mechanism is due to Jurca and Faltings [2008, 2011], who introduce it as building block of a mechanism for opinion polls. My presentation of the 1/prior mechanism is slightly different in that I introduce it as a "correction" of simple output agreement (see Section 2.4). Recall that simple output agreement refers to paying an agent only if her report agrees with that of her peer, and nothing otherwise. For example, simple output agreement could pay \$1 to each agent when both agents report the same signal, and \$0 else. Theorem 2.2 says that simple output agreement is strictly truthful only if, for any possible signal observation, agent i's observed signal is also the most likely signal for her peer agent j, i.e.

$$p(s|s) > p(s'|s) \text{ for all } s, s' \in \{1, \ldots, m\} \text{ with } s' \neq s.$$

But what if one generalizes the notion of output agreement and allows the mechanism to pay different amounts for agreeing on different signals? Intuitively, such a biased output agreement mechanism could reward agreement on unlikely signals more than agreement on likely signals. The 1/prior mechanism uses the signal prior for this purpose.

3.3.1 Mechanism

The *1/prior mechanism* is defined as:

1. Each agent i is asked for her signal report $x_i \in \{1, \ldots, m\}$.

2. For each agent i, choose peer agent $j = i + 1$ (modulo n) and pay agent i:

$$u_i(x_i, x_j) = \begin{cases} \tau \cdot \frac{1}{p(x_i)} & \text{if } x_i = x_j \\ 0 & \text{if } x_i \neq x_j \end{cases}$$

where $\tau > 0$, $p(x_i)$ is the signal prior for signal x_i, and x_j is the signal report by *peer* agent j.

3.3.2 Incentive Analysis

Theorem 3.1. *[Jurca and Faltings, 2011] The 1/prior mechanism is strictly BNIC if and only if $p(s|s) > p(s|s')$ for all signals $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$.*

Proof. Since agent i's utility only depends on her own report and the report of her peer agent j, it is sufficient to consider only these two agents. For all signals $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$, we have:

$$\begin{aligned} & \mathbf{E}_{S_j}\Big[u_i(s, S_j) \mid S_i = s\Big] > \mathbf{E}_{S_j}\Big[u_i(s', S_j) \mid S_i = s\Big] \\ \Leftrightarrow \quad & p(s|s) \cdot \frac{\tau}{p(s)} > p(s'|s) \cdot \frac{\tau}{p(s')} \Leftrightarrow \frac{p(s|s)}{p(s)} > \frac{p(s'|s)}{p(s')} \\ \Leftrightarrow \quad & p(s|s) > p(s|s') \end{aligned}$$

□

3.4 Quadratic Scoring Rule

In contrast to the classical peer prediction method, it is no longer sufficient for the Shadowing Method that the used scoring rule is strictly proper. An important part of the Shadowing Method is the *quadratic scoring rule* [Brier, 1950], which I give in normalized form to yield scores in the interval $[0, 1]$:

$$R_q(y, \omega) = y(\omega) - 0.5 \sum_{k=1}^{m} y(k)^2 + 0.5 \tag{3.1}$$

When the set of possible reports is restricted, properness by itself does not imply that the agent reports the report closest to her true belief. But the quadratic scoring rule R_q does have this property, as I will show.

Definition 11 captures the loss in expected score that an agent incurs by reporting a vector y which may or may not be her true belief b.

Definition 11 (Expected Loss)**.** Let $G(y|b)$ denote the expected score of a scoring rule R given belief $b \in \mathcal{D}$ and report $y \in \mathbb{R}^m$. The *expected loss* $L(y|b)$ of scoring rule R is defined as the amount by which the expected score is less

than that for a truthful report:

$$L(y|b) = G(b|b) - G(y|b).$$

Observe that we do not require $y \in \mathcal{D}$. In particular, the elements of y do not need to sum up to 1, and single elements may be negative or larger than 1. We will need this for the mechanisms in Chapter 5.

This representation of scoring rules as loss functions is due to Savage [1971].

Theorem 3.2 (Quadratic Loss)**.** *[Savage, 1971] The quadratic scoring rule R_q has quadratic expected loss $L_q(y|b) = 0.5 \cdot \sum_{k=1}^{m} \big(b(k) - y(k)\big)^2$.*

Proof. The proof follows Selten [1998, p. 47–48].

For $G_q(y|b)$ we have:

$$\begin{aligned}
G_q(y|b) =& \sum_{k=1}^{m} b(k) \left(y(k) - 0.5 \sum_{o=1}^{m} b(o)^2 + 0.5 \right) \\
=& \sum_{k=1}^{m} b(k)y(k) - 0.5 \sum_{o=1}^{m} y(o)^2 + 0.5 \\
=& \sum_{k=1}^{m} b(k)y(k) - 0.5 \sum_{o=1}^{m} \big(y(o) - b(o)\big)^2 \\
& \qquad + 0.5 \sum_{o=1}^{m} b(o)^2 - 0.5 \sum_{o=1}^{m} 2y(o)b(o) + 0.5 \\
=& 0.5 \sum_{k=1}^{m} b(k)^2 - 0.5 \sum_{k=1}^{m} \big(y(k) - b(k)\big)^2 + 0.5 \\
=& 0.5 \left(\sum_{k=1}^{m} b(k)^2 - \sum_{k=1}^{m} \big(y(k) - b(k)\big)^2 + 1 \right)
\end{aligned}$$

For $L_q(y|b)$ we then have:

$$\begin{aligned}
L_q(y|b) =& G_q(b|b) - G_q(y|b) \\
=& 0.5 \left(\sum_{k=1}^{m} b(k)^2 + 1 \right) - 0.5 \left(\sum_{k=1}^{m} b(k)^2 - \sum_{k=1}^{m} \big(y(k) - b(k)\big)^2 + 1 \right) \\
=& 0.5 \sum_{k=1}^{m} \big(b(k) - y(k)\big)^2
\end{aligned}$$

□

Corollary 3.3 states that when using the quadratic scoring rule, an agent faced with a restricted set of possible belief reports maximizes her expected score by reporting the belief report with minimal Euclidean distance to her true belief. In the words of Friedman [1983], the quadratic scoring rule is *effective* with

respect to the Euclidean distance. Corollary 3.3 follows because maximizing expected score is equivalent to minimizing expected loss.

Corollary 3.3. *Let $b \in \mathcal{D}$ be an agent's true belief about a future event. If the mechanism scores the agent's belief report according to the quadratic scoring rule R_q but restricts the set of allowed reports to $Y \subseteq \mathbb{R}^m$, a rational agent will report $y \in Y$ with minimal $\sum_{k=1}^{m} \big(y(k) - b(k)\big)^2$.*

Corollary 3.4 follows because $L_q(y|b) > 0$ if and only if $y \neq b$.

Corollary 3.4 (Strict Properness)**.** *[Brier, 1950] The quadratic scoring rule R_q is strictly proper.*

3.5 Binary Shadowing Method

The Shadowing Method first takes the known signal prior and, using the agent's signal report, perturbs it into a "shadow posterior," i.e. a belief that is "close" to the agent's true (but unknown) signal posterior. In a second step, this shadow posterior is used as the belief report that is scored using the quadratic scoring rule.

As is usual in peer prediction, the event that is to be predicted is peer agent j's signal report. The Shadowing Method would coincide with the classical peer prediction method presented in Section 2.5 if the mechanism could guarantee that the computed shadow posterior is exactly the agent's true signal posterior. However, since the true signal posterior is no longer assumed to be something the mechanism can compute from a signal report, the mechanism must instead compute a shadow posterior that is close to the signal posterior.

There are two key elements that make the Shadowing Method strictly truthful. First, it has to guarantee that the shadow posterior that is computed based on the true signal is always closer to the true signal posterior than any of the other possible shadow posteriors following dishonest signal reports. Second, the Shadowing Method needs to score the shadow posterior using a scoring rule that is not only strictly proper but satisfies the more demanding property that the closer a belief report is to the true belief, the higher the expected score.

3.5.1 Mechanism

The *Binary Shadowing Method* is defined as:

1. Each agent i is asked for her signal report $x_i \in \{1, 2\}$.

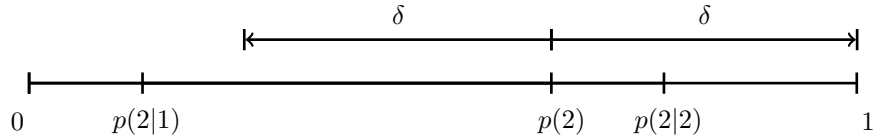


Figure 3.1: Illustration of the Binary Shadowing Method, where $p(2) \in \big(p(2|1), p(2|2)\big)$. Note that $p(2|1)$ is closer to $p'(2|1) = p(2) - \delta$ than to $p'(2|2) = p(2) + \delta$, and that $p(2|2)$ is closer to $p'(2|2) = p(2) + \delta$ than to $p'(2|1) = p(2) - \delta$.

2. Perturb signal prior $p(\cdot)$ using x_i resulting in "shadow posterior" report $p'(\cdot|x_i)$:

$$p'(\cdot|x_i) = \begin{cases} \begin{pmatrix} p(1) + \delta \\ p(2) - \delta \end{pmatrix} & \text{if} \quad x_i = 1 \\ \begin{pmatrix} p(1) - \delta \\ p(2) + \delta \end{pmatrix} & \text{if} \quad x_i = 2, \end{cases} \tag{3.2}$$

   where $\delta = \min\big(p(1), p(2)\big)$.

3. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$, and pay agent i:

$$u_i(x_i, x_j) = R_q\big(p'(\cdot|x_i), x_j\big).$$

   where R_q is the quadratic scoring rule, and x_j is the signal report by *peer* agent j.

3.5.2 Incentive Analysis

Theorem 3.5. *The Binary Shadowing Method is strictly BNIC if and only if* $p(s|s) > p(s)$ *for all* $s \in \{1, 2\}$.

Proof. Suppose agent i's signal is $S_i = 2$, so that her signal posterior is $p(\cdot|2)$. The argument is analogous for signal posterior $p(\cdot|1)$ following $S_i = 1$. The proof is via reasoning about the expected loss $L_q\big(p'(\cdot|x_i)\big|p(\cdot|2)\big)$ between an agent's signal posterior and the shadow posteriors following $x_i = 1$ and $x_i = 2$, respectively. Since there are only two signals, i.e. $m = 2$, minimizing $\sum_{s=1}^{m}\big(p(s|2) - p'(s|x_i)\big)^2$ coincides with minimizing $\big|p(2|2) - p'(2|x_i)\big|$, because whenever the latter is minimal, so is $\big|p(1|2) - p'(1|x_i)\big|$. It is thus sufficient to reason about the distance between $p'(2|x_i)$ and $p(2|2)$, and the Binary Shadowing Method is strictly truthful if and only if $\big|p(2) + \delta - p(2|2)\big| < \big|p(2) - \delta - p(2|2)\big|$. Noting that $\delta > 0$ because $p(1), p(2) > 0$, there are two cases (also compare Figure 3.1):

- $p(2|2) > p(2)$. But now $\delta > 0$ and $p(2) - p(2|2) < 0$, and so $\big|p(2) + \delta - p(2|2)\big| < \big|p(2) - \delta - p(2|2)\big|$.

- $p(2|2) \leq p(2)$. But now $\delta > 0$ and $p(2) - p(2|2) \geq 0$, and so $\big|p(2) + \delta - p(2|2)\big| \geq \big|p(2) - \delta - p(2|2)\big|$. □

Corollary 3.6. *The Binary Shadowing Method is strictly BNIC if and only if* $p(2|2) > p(2) > p(2|1)$.

Proof. If $p(s|s) > p(s)$ for all $s \in \{1, 2\}$, it holds that $p(2|2) > p(2)$ in particular. Similarly, it holds that $p(1|1) > p(1) \Leftrightarrow 1 - p(2|1) > 1 - p(2) \Leftrightarrow p(2) > p(2|1)$. □

Theorem 3.5 means that the Binary Shadowing Method is strictly truthful if the belief that peer agent j observes signal s increases when agent i observes signal s. Another way of saying this is that the Binary Shadowing Method is strictly truthful if signal observations between agents are positively correlated. The proof does not rely on the particular choice of δ but only requires $\delta > 0$.

The Binary Shadowing Method is robust against inaccuracies of the signal prior $p(\cdot)$. In fact, any vector y satisfying $p(s|s) > y(s)$ for all $s \in \{1, 2\}$ can be used in place of the signal prior and the Binary Shadowing Method remains strictly truthful. Moreover, while for the sake of this thesis I use the Shadowing Method exclusively for peer prediction, it can also be used as a proper scoring rule for signals, i.e. with any observable future event ω that does not have to be another agent's signal report.

In Section 3.7.1, we will see that for binary signals, the truthfulness conditions of the classical peer prediction method, the 1/prior mechanism, and the Shadowing Method are equivalent.

3.6 Multi-Signal Shadowing Method

There are two ways to generalize the binary Shadowing Method to $m \geq 3$ signals. The first is a direct generalization, the second reduces the m signal problem to the binary problem. They are equivalent in the sense that they require the same truthfulness condition.

3.6.1 Direct Generalization

Mechanism

The *Multi-Signal Shadowing Method (direct generalization)* is defined as:

1. Each agent i is asked for her signal report $x_i \in \{1, \ldots, m\}$.

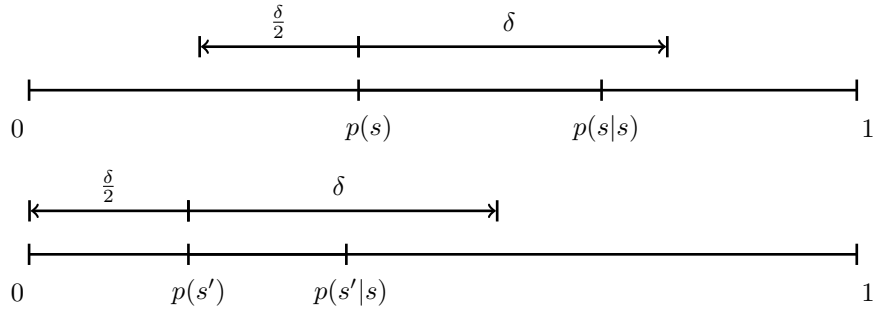


Figure 3.2: Illustration of the Multi-Signal Shadowing Method with $m = 3$, depicting only two signals s and s' with $s' \neq s$. Observe that following $S_i = s$, the belief for both s and s' increases from signal prior to signal posterior. Nevertheless, agent i would maximize her expected score (minimize her expected loss) by reporting $x_i = s$ instead of $x_i = s'$ because the increase is larger for signal s, i.e. $p(s|s) - p(s) > p(s|s') - p(s')$.

2. Perturb the (m-valued) signal prior $p(\cdot)$ using x_i resulting in an (m-valued) "shadow posterior" report (also compare Figure 3.2)

$$p'(\cdot|x_i) = \begin{pmatrix} p(1) - \frac{\delta}{m-1} \\ \cdots \\ p(x_i) + \delta \\ \cdots \\ p(m) - \frac{\delta}{m-1} \end{pmatrix},$$

   where $\delta = \min\Big(\min\limits_{s\in\{1,\ldots,m\}} p(s) \cdot (m-1),\, 1 - \max\limits_{s\in\{1,\ldots,m\}} p(s)\Big)$, so that the entries in $p'(\cdot|x_i)$ are in between 0 and 1.

3. For each agent i, choose peer agent $j = i + 1$ (modulo n), and pay agent i:

$$u_i(x_i, x_j) = R_q\big(p'(\cdot|x_i), x_j\big).$$

   where R_q is the quadratic scoring rule, and x_j is the signal report by *peer* agent j.

Observe that with this choice of δ, the Multi-Signal Shadowing Method coincides with the binary method for $m = 2$.

Incentive Analysis

Theorem 3.7. *The Multi-Signal Shadowing Method (direct generalization) is strictly BNIC if and only if $p(s|s) - p(s) > p(s'|s) - p(s')$ for all $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$.*

Proof. Let agent i observe signal $S_i = s_i$. Compare the truthful report $x_i = s_i$

to the report of some other signal $x_i = s'_i$. The shadow posteriors following these two reports differ only in rows s_i and s'_i. Using R_q's quadratic loss, strict truthfulness is thus equivalent to:

$$
\begin{aligned}
& \mathbf{E}_{S_j}\big[u_i(s_i, S_j)\big] > \mathbf{E}_{S_j}\big[u_i(s'_i, S_j)\big] \\
\Leftrightarrow \quad & L_q\big(p'(\cdot|s_i) \,\big|\, p(\cdot|s_i)\big) < L_q\big(p'(\cdot|s'_i) \,\big|\, p(\cdot|s_i)\big) \\
\Leftrightarrow \quad & \sum_{s=1}^{m}\big(p'(s|s_i) - p(s|s_i)\big)^2 < \sum_{s=1}^{m}\big(p'(s|s'_i) - p(s|s_i)\big)^2 \\
\Leftrightarrow \quad & \Big(p(s_i) + \delta - p(s_i|s_i)\Big)^2 + \Big(p(s'_i) - \frac{\delta}{m-1} - p(s'_i|s_i)\Big)^2 \\
& < \Big(p(s_i) - \frac{\delta}{m-1} - p(s_i|s_i)\Big)^2 + \Big(p(s'_i) + \delta - p(s'_i|s_i)\Big)^2 \\
\Leftrightarrow \quad & p(s_i|s_i) - p(s_i) > p(s'_i|s_i) - p(s'_i)
\end{aligned}
$$

□

Because the quadratic loss is not restricted to reports that are valid probability distributions (Theorem 3.2), Theorem 3.7 holds for any $\delta > 0$, even if it results in a shadow posterior that is not a valid distribution. Moreover, Theorem 3.7 also holds for any $\delta > 0$ if $p'(x_i|x_i) = p(x_i) + \delta$ (as in the above mechanism description) but where all other entries $x'_i \neq x_i$ stay the same as in the signal prior, i.e. $p'(x'_i|x_i) = p(x'_i)$. An advantage of choosing $\delta > 0$ such that $p'(\cdot|x_i)$ is a valid distribution is that the scaling of R_q ensures that ex post payments are in between 0 and 1. For values of $\delta > 0$ that result in $p'(\cdot|x_i)$ that are not valid distributions, R_q will need to be re-scaled appropriately.

Intuitively, the Shadowing Method works by analyzing the belief change from signal prior to signal posterior. More specifically, it assumes that when an agent observes a signal, her belief for another agent observing the same signal increases. With only two possible signal, i.e. $m = 2$, this means that the belief for the signal value that has not been observed has to decrease. For three or more possible signals, however, it can be the case that the beliefs for two or more signals increase, i.e. that the belief for the observed signal is not the only belief that increases. See Figure 3.2 for an illustration. The Multi-Signal Shadowing Method requires that the belief increase for the observed signal is larger than the belief increase for any other signal.

3.6.2 Reduction to Binary Shadowing Method

Mechanism

The *Multi-Signal Shadowing Method (binary method reduction)* is defined as:

1. Each agent i is asked for her signal report $x_i \in \{1, \ldots, m\}$.

2. Choose a signal $k \in \{1, \ldots, m\}$ uniformly at random, and map the signal set $\{1, \ldots, m\}$ to a binary partition $\{\mathbf{1}, \mathbf{2}\}$ by setting

$$\begin{aligned} \mathbf{1} &:= \{1, \ldots, m\} \backslash k \\ \mathbf{2} &:= \{k\}, \end{aligned}$$

with induced binary signal prior:

$$\mathbf{p}(\cdot|k) = \begin{pmatrix} \mathbf{p}(\mathbf{1}|k) \\ \mathbf{p}(\mathbf{2}|k) \end{pmatrix} = \begin{pmatrix} 1 - p(k) \\ p(k) \end{pmatrix},$$

where, with $S_j = s_j$, signal $\mathbf{1}$ occurs when $s_j \in \mathbf{1}$ and signal $\mathbf{2}$ when $s_j \in \mathbf{2}$.

3. Proceed with the induced binary signal prior as in the binary Shadowing Method: perturb $\mathbf{p}(\cdot|k)$ using x_i resulting in binary "shadow posterior" report $\mathbf{p}'(\cdot|x_i, k)$:

$$\mathbf{p}'(\cdot|x_i, k) = \begin{cases} \begin{pmatrix} \mathbf{p}(\mathbf{1}|k) + \delta \\ \mathbf{p}(\mathbf{2}|k) - \delta \end{pmatrix} = \begin{pmatrix} 1 - p(k) + \delta \\ p(k) - \delta \end{pmatrix} & \text{if} \quad x_i \neq k \\ \begin{pmatrix} \mathbf{p}(\mathbf{1}|k) - \delta \\ \mathbf{p}(\mathbf{2}|k) + \delta \end{pmatrix} = \begin{pmatrix} 1 - p(k) - \delta \\ p(k) + \delta \end{pmatrix} & \text{if} \quad x_i = k, \end{cases} \tag{3.3}$$

where $\delta = \min\big(p(1), \ldots, p(m)\big)$.

4. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$, and pay agent i:

$$u_i(x_i, x_j) = R_q\big(\mathbf{p}'(\cdot|x_i, k), x_j\big).$$

where R_q is the quadratic scoring rule, and x_j is the signal report by peer agent j.

While this is a randomized mechanism, it can be transformed into a deterministic mechanism by averaging over all possible k's on behalf of the agent. The resulting payment of the deterministic mechanism is

$$u_i(x_i, x_j) = \sum_{k=1}^{m} \frac{1}{m} R_q\big(\mathbf{p}'(\cdot|x_i, k), x_j\big).$$

A very similar idea of de-randomization has been proposed by Matheson and Winkler [1976, p. 1090] in the context of designing proper scoring rules for continuous variables that build upon proper scoring rules for binary variables.

Incentive Analysis

Theorem 3.8. *The Multi-Signal Shadowing Method (binary method reduction) is strictly BNIC if and only if* $p(s|s) - p(s) > p(s'|s) - p(s')$ *for all* $s, s' \in \{1, \dots, m\}$ *with* $s' \neq s$.

Proof. WLOG, let agent i observe signal $S_i = s$. Furthermore, let

$$\mathbf{p}(\cdot|s,k) = \begin{pmatrix} \mathbf{p}(\mathbf{1}|s,k) \\ \mathbf{p}(\mathbf{2}|s,k) \end{pmatrix} = \begin{pmatrix} 1 - p(k|s) \\ p(k|s) \end{pmatrix}$$

denote agent i's induced (binary) signal posterior, and let

$$L_q\big(\mathbf{p}'(\cdot|x_i,k)\big|\mathbf{p}(\cdot|s,k)\big) = \big(\mathbf{p}'(\mathbf{1}|x_i,k) - \mathbf{p}(\mathbf{1}|s,k)\big)^2 + \big(\mathbf{p}'(\mathbf{2}|x_i,k) - \mathbf{p}(\mathbf{2}|s,k)\big)^2$$

denote the expected loss of being scored using induced (binary) shadow posterior $\mathbf{p}'(\cdot|x_i,k)$ instead of the true induced (binary) signal posterior $\mathbf{p}(\cdot|s,k)$, and given that the mechanism chose the random partition associated with signal k.

Compare the truthful report $x_i = s$ to the report of some other signal $x_i = s'$. The scores of these two reports differ only if $k = s$ or $k = s'$, and since k is chosen uniformly, these two cases are equally likely. We thus have:

$$\begin{aligned}
& \mathbf{E}_{S_j,k}\big[u_i(s,S_j)\big] > \mathbf{E}_{S_j,k}\big[u_i(s',S_j)\big] \\
\Leftrightarrow\quad & L_q\big(\mathbf{p}'(\cdot|s,k=s)\big|\mathbf{p}(\cdot|s,k=s)\big) + L_q\big(\mathbf{p}'(\cdot|s,k=s')\big|\mathbf{p}(\cdot|s,k=s')\big) \\
< {} & L_q\big(\mathbf{p}'(\cdot|s',k=s)\big|\mathbf{p}(\cdot|s,k=s)\big) + L_q\big(\mathbf{p}'(\cdot|s',k=s')\big|\mathbf{p}(\cdot|s,k=s')\big) \\
\Leftrightarrow\quad & \big(\mathbf{p}'(\mathbf{1}|s,k=s) - \mathbf{p}(\mathbf{1}|s,k=s)\big)^2 + \big(\mathbf{p}'(\mathbf{2}|s,k=s) - \mathbf{p}(\mathbf{2}|s,k=s)\big)^2 \\
& + \big(\mathbf{p}'(\mathbf{1}|s,k=s') - \mathbf{p}(\mathbf{1}|s,k=s')\big)^2 + \big(\mathbf{p}'(\mathbf{2}|s,k=s') - \mathbf{p}(\mathbf{2}|s,k=s')\big)^2 \\
< {} & \big(\mathbf{p}'(\mathbf{1}|s',k=s) - \mathbf{p}(\mathbf{1}|s,k=s)\big)^2 + \big(\mathbf{p}'(\mathbf{2}|s',k=s) - \mathbf{p}(\mathbf{2}|s,k=s)\big)^2 \\
& + \big(\mathbf{p}'(\mathbf{1}|s',k=s') - \mathbf{p}(\mathbf{1}|s,k=s')\big)^2 + \big(\mathbf{p}'(\mathbf{2}|s',k=s') - \mathbf{p}(\mathbf{2}|s,k=s')\big)^2 \\
\Leftrightarrow\quad & \Big(1 - p(s) - \delta - \big(1 - p(s|s)\big)\Big)^2 + \Big(p(s) + \delta - p(s|s)\Big)^2 \\
& + \Big(1 - p(s') + \delta - \big(1 - p(s'|s)\big)\Big)^2 + \Big(p(s') - \delta - p(s'|s)\Big)^2 \\
< {} & \Big(1 - p(s) + \delta - \big(1 - p(s|s)\big)\Big)^2 + \Big(p(s) - \delta - p(s|s)\Big)^2 \\
& + \Big(1 - p(s') - \delta - \big(1 - p(s'|s)\big)\Big)^2 + \Big(p(s') + \delta - p(s'|s)\Big)^2 \\
\Leftrightarrow\quad & \delta\big(p(s) - p(s') - p(s|s) + p(s'|s)\big) < 0 \\
\Leftrightarrow\quad & p(s|s) - p(s) > p(s'|s) - p(s')
\end{aligned}$$

□

Mechanism	Condition required for strict truthfulness
Classical Peer Prediction	$p(\cdot\|s) \neq p(\cdot\|s')$
1/prior	$p(s\|s) > p(s\|s')$
Shadowing Method	$p(s\|s) - p(s) > p(s'\|s) - p(s')$

Table 3.1: The conditions are for all $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$.

3.7 Comparison of Mechanisms

To ensure strict truthfulness, peer prediction mechanisms require certain truthfulness conditions that restrict the allowed belief models in some way. For example, the classical peer prediction method requires stochastic relevance, i.e. that there is a one to one mapping from signals to signal posteriors. Table 3.1 provides a summary of conditions for the mechanisms we have discussed. This section is devoted to analyzing their relationship with one another. Section 3.7.1 provides a comparison of the conditions required with a binary signal set. Section 3.7.2 discusses the relationship between the conditions with three or more possible signals.

3.7.1 Binary Signals

Lemma 3.9. *With binary signals, and given the world state model as described in Section 2.1, stochastic relevance implies $p(s|s) > p(s)$ for all $s \in \{1, 2\}$.*

Proof. I show that $p(2|2) > p(2)$ given stochastic relevance and the world state model. The case for $p(1|1) > p(1)$ is analogous. The proof proceeds in three steps:

First, associate every state with one of two groups H and L. Associate states $t \in \{1, \ldots, l\}$ for which $\Pr(S = 2|T = t) > p(2) = \Pr(S = 2)$ with group H, and states $t \in \{1, \ldots, l\}$ for which $\Pr(S = 2|T = t) \leq p(2)$ with group L. That is, the states in group H are those that put more weight on signal 2 than the signal prior, and the states in group L are those that put less or equal weight on signal 2 than the signal prior.

Second, both H and L are non-empty, i.e. there are states $t, t' \in \{1, \ldots, l\}$, such that $\Pr(S = 2|T = t) > \Pr(S = 2)$ and $\Pr(S = 2|T = t') \leq \Pr(S = 2)$.

This is the case because:

1. From the definition of the signal posterior (Equation 2.2), we know that for all $s \in \{1, 2\}$:

$$p(2|s) = \sum_{t=1}^{l} \Pr(S_j = 2 \mid T = t) \cdot \Pr(T = t \mid S_i = s).$$

Stochastic relevance means that $p(2|2) \neq p(2|1)$, so that $\Pr(S = 2|T = t) \neq \Pr(S = 2|T = t')$ for some states $t, t' \in \{1, \ldots, l\}$.

2. It cannot be that $\Pr(S = 2|T = t) \leq \Pr(S = 2)$ for all $t \in \{1, \ldots, l\}$ or $\Pr(S = 2|T = t) > \Pr(S = 2)$ for all $t \in \{1, \ldots, l\}$ because (Equation 2.4):

$$p(2) = \Pr(S = 2) = \sum_{t=1}^{l} \Pr(S = 2|T = t) \cdot \Pr(T = t).$$

That is, the signal prior is the *average* of signal conditionals $\Pr(S = 2|T = t)$ weighted by the state belief $\Pr(T = t)$. The equality case $\Pr(S = 2|T = t) = \Pr(S = 2)$ for all t is excluded since that would imply $p(2|2) = p(2)$ and by $p(2) = p(2|1)p(1) + p(2|2)p(2)$ also $p(2|1) = p(2)$, and a contradiction to stochastic relevance.

Moreover, both groups have positive probability because the world state model from Section 2.1 demands that $\Pr(T = t) > 0$ for all $t \in \{1, \ldots, l\}$.

Third, the probability of states in group H increases given observation $S = 2$. To see this, we need to know for which $t \in \{1, \ldots, l\}$ is $\Pr(T = t|S = 2) > \Pr(T = t)$ and obtain:

$$\begin{aligned} &\Pr(T = t|S = 2) > \Pr(T = t) \\ \Leftrightarrow &\frac{\Pr(S = 2|T = t) \cdot \Pr(T = t)}{\Pr(S = 2)} > \Pr(T = t) \\ \Leftrightarrow &\Pr(S = 2|T = t) > \Pr(S = 2). \end{aligned}$$

That is, exactly those states in group H become more likely after signal 2. The statement follows because states in group H have more weight on signal 2 than the signal prior $p(2)$ and become more likely after $S = 2$. □

Theorem 3.10. *With binary signals, the classical peer prediction method, the 1/prior mechanism, and the Shadowing Method are strictly truthful in the same settings.*

Proof. It needs to be shown that the following conditions are equivalent for all $s', s \in \{1, 2\}$ with $s' \neq s$:

- $p(\cdot|s) \neq p(\cdot|s')$ (stochastic relevance).
- $p(s|s) > p(s|s')$.
- $p(s|s) > p(s)$.

Lemma 3.9 shows that $p(\cdot|s) \neq p(\cdot|s')$ and $p(s|s) > p(s)$ are equivalent. Moreover,

$$p(s|s') = \frac{p(s'|s)\,p(s)}{p(s')} = \frac{\big(1 - p(s|s)\big)\,p(s)}{1 - p(s)},$$

and so $p(s|s) > p(s|s') \Leftrightarrow p(s|s)\big(1 - p(s)\big) > p(s)\big(1 - p(s|s)\big) \Leftrightarrow p(s|s) > p(s)$. □

3.7.2 More than Two Signals

With only two signals, a belief increase in signal $s \in \{1, 2\}$ means that the belief for signal $s' \neq s$ decreases. For three or more possible signals, however, it can be the case that, following signal observation s, the beliefs for several signals increase. The Multi-Signal Shadowing Method requires that the belief increase for the observed signal is larger than the belief increase for any other signal:

$$\begin{aligned} p(s|s) - p(s) &> p(s'|s) - p(s') \\ \text{for all } \; s, s' &\in \{1, \ldots, m\} \;\text{ with }\; s' \neq s. \end{aligned} \tag{3.4}$$

The condition of the 1/prior mechanism can also be written in terms of a belief change from signal prior to signal posterior using Bayes' law:

$$\begin{aligned} p(s|s) > p(s|s') \Leftrightarrow \frac{p(s|s)}{p(s)} &> \frac{p(s'|s)}{p(s')} \\ \text{for all } \; s, s' &\in \{1, \ldots, m\} \;\text{ with }\; s' \neq s. \end{aligned} \tag{3.5}$$

This condition also means that the belief increase for the observed signal is larger than the belief increase for any other signal. The difference between the two conditions is what "largest" means: the *absolute* change (Shadowing Method) or the *relative* change (1/prior mechanism). These two conditions are incomparable, i.e. there are settings where the first condition holds but the second doesn't, and vice versa.

Theorem 3.11. *With more than two signals, the conditions required by the 1/prior mechanism and the Shadowing Method are incomparable.*

Proof. The proof is by example. An example of beliefs, where the Shadowing Method is strictly truthful and the 1/prior is not, is the following with $m = 3$

signals:

$$\Pr(T) = \begin{pmatrix} 0.25 \\ 0.5 \\ 0.25 \end{pmatrix}$$

$$\Pr(S|T) = \begin{pmatrix} 0.15 & 0.5 & 0.6 \\ 0.05 & 0.3 & 0.35 \\ 0.35 & 0.2 & 0.05 \end{pmatrix}.$$

This results in signal prior

$$p(\cdot) = \Pr(S|T) \times \Pr(T) = \begin{pmatrix} 0.475 \\ 0.3625 \\ 0.2 \end{pmatrix}$$

and signal posterior matrix

$$p(\cdot|\cdot) = \begin{pmatrix} 0.50428571 & 0.40344828 & 0.353125 \\ 0.33428571 & 0.38103448 & 0.390625 \\ 0.16142857 & 0.21551724 & 0.25625 \end{pmatrix}$$

which satisfies $p(s|s) - p(s) > p(s'|s) - p(s')$ but not $p(s|s) > p(s|s')$ for all $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$ since $p(2|2) = 0.38103448 \not> 0.390625 = p(2|3)$.

Similarly, an example of beliefs, where the 1/prior mechanism is strictly truthful but where the Shadowing Method is not is the following with $m = 3$ signals:

$$\Pr(T) = \begin{pmatrix} 0.52951336 \\ 0.22410641 \\ 0.24638023 \end{pmatrix}$$

$$\Pr(S|T) = \begin{pmatrix} 0.44778266 & 0.74312272 & 0.07829395 \\ 0.34278219 & 0.20756903 & 0.74654266 \\ 0.20943515 & 0.04930825 & 0.1751634 \end{pmatrix}.$$

This results in signal prior

$$p = \begin{pmatrix} 0.42293554 \\ 0.41195865 \\ 0.1651058 \end{pmatrix}$$

and signal posterior matrix

$$p(\cdot|\cdot) = \begin{pmatrix} 0.54722581 & 0.31616071 & 0.37096916 \\ 0.30795506 & 0.50778717 & 0.43927101 \\ 0.14481914 & 0.17605212 & 0.18975984 \end{pmatrix},$$

which satisfies $p(s|s) > p(s|s')$ but not $p(s|s) - p(s) > p(s'|s) - p(s')$ for all $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$ since $p(3|3) - p(3) = 0.18975984 - 0.1651058 = 0.02465404 \not> 0.02731236 = 0.43927101 - 0.41195865 = p(2|3) - p(2)$. □

In contrast to binary settings, the condition required by the classical peer prediction method is weaker than the conditions of both the 1/prior mechanism and the Shadowing Method.

Theorem 3.12. *The conditions required by the 1/prior mechanism and the Shadowing Method are strictly stronger than the condition required by the classical peer prediction method.*

Proof. It needs to be shown that any setting satisfying either the condition of the 1/prior mechanism or the condition of the Shadowing Method, also satisfies stochastic relevance. I show the equivalent statement that whenever stochastic relevance does not hold, neither does the condition of the 1/prior mechanism nor that of the Shadowing Method. If stochastic relevance does not hold, it must be that $p(\cdot|s) = p(\cdot|s')$ for some $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$. From that, it follows that $p(s|s) = p(s|s')$, so that the 1/prior condition is violated. Furthermore, it follows that $p(s'|s') = p(s'|s)$. Now assume that $p(s|s) - p(s) > p(s'|s) - p(s')$, but then $p(s'|s') - p(s') = p(s'|s) - p(s') < p(s|s) - p(s) = p(s|s') - p(s)$. Moreover, the containment is strict because both example settings in the proof of Theorem 3.11 satisfy stochastic relevance. □

3.8 Conclusion

The truthfulness condition of the 1/prior mechanism and the truthfulness condition of the Shadowing Method are incomparable, in that there exist belief settings, where the 1/prior mechanism is strictly truthful but the Shadowing Method is not, and vice versa. There are, however, compelling reasons for choosing the Shadowing Method over the 1/prior mechanism. In particular, the Shadowing Method allows the designer to bound the range of the (ex post) payments without needing to know the signal prior. This becomes important when using the methods as building blocks for other mechanisms. This difference is exemplified in Chapter 4, where both the Shadowing Method and the 1/prior

mechanism are used to design Bayesian Truth Serum mechanisms, which work without the mechanism knowing anything about the belief model.

An interesting direction for future work is to study the properties of Shadowing Methods based on metrics other than the Euclidean metric. To achieve strict truthfulness, this would need to be paired with other scoring rules that are "effective" with respect to these metrics. A natural candidate is the spherical rule that is known to be effective with respect to the renormalized Euclidean metric [Friedman, 1983].

Chapter 4

The Robust Bayesian Truth Serum

The classical peer prediction method (Chapter 2) is truthful for any finite number of agents $n \geq 2$ but relies on a common belief model, shared by all agents and the mechanism. The Shadowing Method (Chapter 3) relaxes this assumption in that the agents share a common belief model but the mechanism only needs to know the signal prior. The Bayesian Truth Serum (BTS) by Prelec [2004] further relaxes the common knowledge required by the mechanism. While BTS still assumes that the agents share a common belief model, this model need not be known by the mechanism—not even the signal prior.

In addition to the signal report, BTS also asks each agent for a *prediction report*, which reflects the agent's belief about the distribution of signals in the population. An agent's payment depends on both reports, with a signal component that rewards reports that are "surprisingly common," i.e., more common than collectively predicted, and a prediction component that rewards accurate predictions of the reports made by others.

In addition to requiring an additional report, a significant drawback of BTS when compared to the mechanisms presented in Chapters 2 and 3 is that it only aligns incentives for a large enough number of agents, where this number depends on the belief model, which is assumed unknown to the mechanism. In addition, BTS may leave a participant with a negative payment, and is not numerically robust for all inputs.

In this chapter, I present a Robust Bayesian Truth Serum (RBTS) that is strictly truthful for any number of agents $n \geq 3$ (for more than two signals) and any number of agents $n \geq 2$ (for binary signals). RBTS uses the Shadowing Method from Chapter 3 as a building block and is the first peer prediction

mechanism that does not rely on knowledge of the common belief model to provide strict incentive compatibility for any number of agents $n \geq 3$. RBTS takes the same reports as BTS, and an agent's payment continues to consist of one component that depends on an agent's signal report and a second component that depends on an agent's prediction report. In contrast to the original BTS, RBTS is ex post individually rational (so that no agent makes a negative payment in any outcome) and numerically robust (well defined for all possible reports). Moreover, RBTS seems conceptually simpler than BTS, and the incentive analysis is more straightforward.

I also present the 1/posterior BTS [Radanovic and Faltings, 2013], which was published after RBTS, and which is strictly truthful for any $n \geq 2$ and multiple signals. When first published, RBTS was only defined for binary signals but in this thesis it is extended to multiple signals. I compare the truthfulness conditions of 1/posterior BTS and RBTS in Section 4.7 and show that they are identical for binary-signal settings and incomparable for more than two signals. While RBTS still requires $n \geq 3$ agents for more than two signals, it has the important property that the designer can set an upper and a lower bound on the ex post payments. In contrast, the 1/posterior BTS may end up having to pay agents effectively unbounded amounts.

4.1 Related Work

In addition to the classical peer prediction method, the Shadowing Method, and the original BTS, there is other related work.

Jurca and Faltings [2007] extend the original peer prediction method to allow agents to have small deviations from a common belief model, that is known to the mechanism. They establish a trade-off between the expected payment of the mechanism and the robustness to deviations from the model. In comparison, BTS schemes do not assume any knowledge about the common belief model on behalf of the mechanism.

Jurca and Faltings [2008] assume a common belief model known to the agents but unknown to the mechanism in a on-line polling setting, where the current empirical frequency of signal reports is published and updated as agents arrive. While their mechanism only requires a signal report (and not a prediction report), it is not incentive compatible. Rather, the authors show that—when agents are strategic—the signal reports converge towards the true distribution of signals in the population. Moreover, one of their main criticisms of BTS is that it needs to withhold all information reports until the end of the poll. This criticism does not apply to RBTS, which easily adapts to online settings by sequentially scoring groups of three agents, and subsequently releasing their

reports (which can be published as empirical frequencies). See Chapter 6 for a more detailed discussion of this work.

A setting similar to on-line polling is studied by Lambert and Shoham [2008], and in this case without requiring a common belief model to agents. However, their mechanism is only *weakly* incentive compatible, i.e., in the equilibrium, agents are indifferent between being truthful and misreporting. For this reason it does not extend to settings in which providing accurate information is costly or when agents have some other outside incentive for making false reports.

4.2 Model

The model adopted in this chapter is the variation of the standard model (Section 2.1), where the belief model is common knowledge amongst all agents, but the mechanism does not need to know it. Moreover, it is assumed that the signal posteriors are fully mixed, i.e. $p(s|s') > 0$ for all $s, s' \in \{1, \ldots, m\}$. In terms of basic model parameters, this condition is, for example, satisfied if $\Pr(S = s|T = t) > 0$ for all $s \in \{1, \ldots, m\}$ and $t \in \{1, \ldots, l\}$.

4.3 Bayesian Truth Serum (BTS)

In this section, I explain the original Bayesian Truth Serum (BTS) by Prelec [2004]. Prelec presents two versions of BTS, one for an infinite number of agents $n \to \infty$ and one for finite n with $n \geq 3$. Given the focus of my work, I present the latter version. While I present the binary version of this mechanism, BTS is also defined for an arbitrary number of signals.

4.3.1 Mechanism

The *original Bayesian Truth Serum (BTS)* with binary signals and finite populations is defined as:

1. Each agent i is asked for two reports:

   - **Signal report**: Let $x_i \in \{1, 2\}$ be agent i's reported signal.
   - **Prediction report**: Let $y_i \in \mathcal{D}$ be agent i's reported signal posterior.

2. Let $x_k(s)$ be the function that returns 1 if $x_k = s$ and 0 otherwise. Then, for both possible signals and for each agent $j \neq i$, calculate the empirical frequency[1] of all agents' signal reports except those of agents i and j:

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

$$\bar{x}_{-ij}(1) = \frac{1}{n}\left(\left(\sum_{k\neq i,j} x_k(1)\right) + 1\right), \qquad \bar{x}_{-ij}(2) = 1 - \bar{x}_{-ij}(1)$$

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

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

4. Pay agent i:

$$u_i(x_i, y_i, x_{-i}, y_{-i}) = \underbrace{\sum_{j\neq i}\left(x_i(1)\ln\left(\frac{\bar{x}_{-ij}(1)}{\bar{y}_{-ij}(1)}\right) + x_i(2)\ln\left(\frac{\bar{x}_{-ij}(2)}{\bar{y}_{-ij}(2)}\right)\right)}_{\text{signal score}} + \underbrace{\sum_{j\neq i}\left(\bar{x}_{-ij}(1)\ln\left(\frac{y_i(1)}{\bar{x}_{-ij}(1)}\right) + \bar{x}_{-ij}(2)\ln\left(\frac{y_i(2)}{\bar{x}_{-ij}(2)}\right)\right)}_{\text{prediction score}}$$

For the incentive analysis, the equation for u_i can be simplified by replacing the summations over $j \neq i$ with the signal and prediction scores computed using just one, randomly selected, $j \neq i$.

4.3.2 Analysis

First note that the game-theoretic concepts introduced in Section 2.3, including strategies and equilibrium concepts, generalize to mechanisms with both signal and belief reports in the natural way.

Theorem 4.1. *Prelec [2004] The original Bayesian Truth Serum is strictly Bayes-Nash incentive compatible for $n \to \infty$ given the belief model satisfies stochastic relevance.*

The intuition for this result is the following. As Prelec points out, the prediction score is the relative entropy [Kullback and Leibler, 1951] between the empirical distribution of signals and the prediction of that distribution. In particular, the prediction score is maximized when the prediction of the signal distribution exactly matches the empirical frequency of signals. As we will see later in this chapter, another way to achieve strict truthfulness for the prediction report is to score it with a strictly proper scoring rule (Section 2.5.1) applied to the signal report of a single, randomly selected peer agent.

To get an intuition for the signal score, first note that x_i selects whether the left or the right part of the signal score term is selected. Now assume agent i observed signal $S_i = s$. This is a piece of information that she knows but that others do not know, and so she believes that the empirical frequency of signal s will be high relative to the collective prediction for signal s. To describe this intuition, Prelec says that agent i expects that signal s will be "surprisingly common" [Prelec, 2004, p. 462].

Prelec also suggests, but without offering a proof, that the result holds for suitably large, finite n with the required population size depending on agents' belief model. In any case, a challenge with this claim is that the number of agents required depends on the model, which is assumed to be unknown to the designer. Another problem is that the BTS need not satisfy interim individual rationality for small groups, meaning that an agent's expected payment conditioned on knowledge of her own signal may be negative.

Theorem 4.2. *The original Bayesian Truth Serum is not Bayes-Nash incentive compatible or interim IR for $n = 3$ even if the belief model satisfies stochastic relevance.*

This limitation of BTS can be understood from Prelec's treatment of BTS.

Generally, the number of agents required for BTS to be Bayes-Nash incentive compatible depends on the belief model and is hard to characterize. Still, BTS has been discussed in various places without noting this important caveat [e.g. Jurca and Faltings, 2008; Chen and Pennock, 2010]. For this reason, I provide a concrete example. The example is not unique, and does not rely on $n = 3$.

Example 4 (BTS and $n = 3$)**.** *Consider three agents sharing the belief model from Example 1 on p. 15 with $m = 2$ possible signals, where the signal posteriors are given by $p(2|2) = 0.46$ and $p(2|1) = 0.18$. Note that whenever needed, I will use rounded numbers.*

Consider agent $i = 1$, and assume agents 2 and 3 are truthful. Assume that $S_1 = 1$, so that agent 1's truthful reports are $x_1 = 1$ and

$$y_1 = \begin{pmatrix} y_1(1) \\ y_1(2) \end{pmatrix} = \begin{pmatrix} p(1|1) \\ p(2|1) \end{pmatrix} = \begin{pmatrix} 0.82 \\ 0.18 \end{pmatrix}.$$

The expected score for the terms that corresponds to agent $j = 2$ when agent 1 reports truthfully is:

$$\mathbf{E}\left[\ln\left(\frac{\bar{X}_{-12}(1)}{\bar{Y}_{-12}(1)}\right) + \bar{X}_{-12}(1)\ln\left(\frac{0.82}{\bar{X}_{-12}(1)}\right) + \bar{X}_{-12}(2)\ln\left(\frac{0.18}{\bar{X}_{-12}(2)}\right) \Big| S_1 = 1\right],$$

where the expectation is taken with respect to the random variables $\bar{X}_{-12}(1)$, $\bar{X}_{-12}(2) = 1 - \bar{X}_{-12}(1)$, *and* $\bar{Y}_{-12}(1)$. *With probability* $p(1|1) = 0.82$, *agent 1 believes that agent 3 ("agent* $-ij$*") received signal* $S_3 = 1$, *so that*

$$\bar{X}_{-12}(1) = \frac{1}{3}(1+1) = \frac{2}{3} \quad \textit{and} \quad \bar{Y}_{-12}(1) = p(1|1) = 0.82,$$

and with probability $p(2|1) = 0.18$ *that* $S_3 = 2$, *so that*

$$\bar{X}_{-12}(1) = \frac{1}{3}(0+1) = \frac{1}{3} \quad \textit{and} \quad \bar{Y}_{-12}(1) = p(1|2) = 0.54.$$

Given this, we have expected signal score

$$\mathbf{E}\left[\ln\left(\frac{\bar{X}_{-12}(1)}{\bar{Y}_{-12}(1)}\right) \Big| \, S_1 = 1\right] = 0.82 \, \ln\left(\frac{2/3}{0.82}\right) + 0.18 \, \ln\left(\frac{1/3}{0.54}\right) = -0.257$$

and expected prediction score

$$\begin{aligned}
&\mathbf{E}\left[\bar{X}_{-12}(1)\ln\left(\frac{0.82}{\bar{X}_{-12}(1)}\right) + \bar{X}_{-12}(2)\ln\left(\frac{0.18}{\bar{X}_{-12}(2)}\right) \Big| \, S_1 = 1\right] \\
= \quad & 0.82\left(\frac{2}{3}\ln\left(\frac{0.82}{2/3}\right) + \frac{1}{3}\ln\left(\frac{0.18}{1/3}\right)\right) + 0.18\left(\frac{1}{3}\ln\left(\frac{0.82}{1/3}\right) + \frac{2}{3}\ln\left(\frac{0.18}{2/3}\right)\right) \\
= & -0.158
\end{aligned}$$

giving an expected score of $-0.257 - 0.158 = -0.415$ *for* $j = 2$. *Considering also the score due to* $j = 3$, *the total expected score when agent 1 is truthful is* $-0.415 - 0.415 = -0.83$. *BTS fails interim IR.*

Imagine now that agent 1 misreports her signal, i.e. $x_1 = 2$, *while still reporting her prediction report truthfully, i.e.* $y_1 = p(\cdot|1)$. *The expected signal score component for the* $j = 2$ *terms would then become*

$$\mathbf{E}\left[\ln\left(\frac{\bar{X}_{-12}(2)}{\bar{Y}_{-12}(2)}\right) \Big| \, S_1 = 1\right] = 0.82 \, \ln\left(\frac{1/3}{0.18}\right) + 0.18 \, \ln\left(\frac{2/3}{0.46}\right) = 0.572$$

which combines with the prediction score to give 0.414, *and thus, considering also* $j = 3$, *this yields a total expected score of* 0.828. *Agent 1 can do better by making a misreport.*

Example 5 (BTS and $n \to \infty$)**.** *Consider the same belief model as in Example 4 but now a large number of agents. In the limit, and with respect to the beliefs of agent 1 following* $S_1 = 1$, *random variables* $\overline{X}_{-ij}(1)$, $\overline{X}_{-ij}(2)$, $\overline{Y}_{-ij}(1)$ *and*

$\bar{Y}_{-ij}(2)$ *take on their respective values with probability 1:*

$$\begin{aligned}
\mathbf{E}\Big[\overline{X}_{-1j}(1)\,\Big|\,S_1=1\Big] &= \lim_{n\to\infty}\frac{1}{n}\big((n-2)p(1|1)+1\big) = p(1|1) = 0.82\\
\mathbf{E}\Big[\overline{X}_{-1j}(2)\,\Big|\,S_1=1\Big] &= 1-\overline{X}_{-ij}(1) = p(2|1) = 0.18\\
\mathbf{E}\Big[\overline{Y}_{-1j}(1)\,\Big|\,S_1=1\Big] &= \lim_{n\to\infty}\left(\Big(p(1|1)^{(n-2)\cdot p(1|1)}\Big)\Big(p(1|2)^{(n-2)\cdot p(2|1)}\Big)\right)^{1/(n-2)}\\
&= p(1|1)^{p(1|1)}\cdot p(1|2)^{p(2|1)} = 0.82^{0.82}\cdot 0.54^{0.18} = 0.76\\
\mathbf{E}\Big[\bar{Y}_{-1j}(2)\,\Big|\,S_1=1\Big] &= p(2|1)^{p(1|1)}\cdot p(2|2)^{p(2|1)} = 0.18^{0.82}\cdot 0.46^{0.18} = 0.213
\end{aligned}$$

If agent 1 reports truthfully, i.e. $x_1 = 1$ *and* $y_1 = p(\cdot|1)$*, her expected signal score is*

$$\mathbf{E}\left[\ln\left(\frac{\bar{X}_{-1j}(1)}{\bar{Y}_{-1j}(1)}\right)\,\Big|\,S_1=1\right] = \ln\left(\frac{0.82}{0.76}\right) = 0.076$$

and her expected prediction score is

$$\begin{aligned}
&\mathbf{E}\left[\bar{X}_{-1j}(1)\ln\left(\frac{0.82}{\bar{X}_{-1j}(1)}\right)+\bar{X}_{-1j}(2)\ln\left(\frac{0.18}{\bar{X}_{-1j}(2)}\right)\,\Big|\,S_1=1\right]\\
=&\,0.82\ln\left(\frac{0.82}{0.82}\right)+0.18\ln\left(\frac{0.18}{0.18}\right) = 0,
\end{aligned}$$

i.e. 0.076 *in total. A misreport of* $x_1 = 2$ *gives expected signal score (and thus total score) of*

$$\mathbf{E}\left[\ln\left(\frac{\bar{X}_{-1j}(2)}{\bar{Y}_{-1j}(2)}\right)\,\Big|\,S_1=1\right] = \ln\left(\frac{0.18}{0.213}\right) = -0.168.$$

BTS is Bayes-Nash incentive compatible in the large n *limit in the example.*

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

Proposition 4.3. *The score in the original Bayesian Truth Serum is unboundedly negative for prediction reports* y_i *with* $y_i(s) = 0$ *for any* $s \in \{1, 2\}$*.*

4.4 1/posterior Bayesian Truth Serum

The 1/posterior BTS mechanism [Radanovic and Faltings, 2013] uses the technique from the 1/prior mechanism (Section 3.3). It improves upon the original BTS mechanism in that it requires only $n \geq 2$ agents for strict truthfulness.

4.4.1 Mechanism

The *1/posterior Bayesian Truth Serum* is defined as:

1. Each agent i is asked for two reports:
   - **Signal report**: Let $x_i \in \{1, \dots, m\}$ be agent i's reported signal.
   - **Prediction report**: Let $y_i \in \mathcal{D}$ be agent i's reported signal posterior.
2. For each agent i, choose peer agent $j = i + 1$ (modulo n), and pay agent i:

$$u_i(x_i, x_j, y_j) = \begin{cases} \tau \cdot \frac{1}{y_j(x_i)} & \text{if} \quad x_i = x_j \\ 0 & \text{else} \end{cases},$$

   where $\tau > 0$, x_j is the signal report by peer agent j, and y_j is the prediction report of peer agent j.

4.4.2 Incentive Analysis

Theorem 4.4. *[Radanovic and Faltings, 2013] The 1/posterior Bayesian Truth Serum is strictly BNIC if and only if $p(s|s) > p(s|s')$ for all signals $s, s' \in \{1, \dots, m\}$ with $s' \neq s$.*

Proof. Since agent i's utility only depends on her own report and the report of her peer agent j, it is sufficient to consider only these two agents. For all signals $s, s' \in \{1, \dots, m\}$ with $s' \neq s$, we have

$$\begin{aligned} & \mathbf{E}_{S_j}\Big[u_i\big(s, S_j, p(\cdot|S_j)\big) \mid S_i = s\Big] > \mathbf{E}_{S_j}\Big[u_i\big(s', S_j, p(\cdot|S_j)\big) \mid S_i = s\Big] \\ \Leftrightarrow \quad & p(s|s) \cdot \frac{\tau}{p(s|s)} > p(s'|s) \cdot \frac{\tau}{p(s'|s')} \Leftrightarrow 1 > \frac{p(s'|s)}{p(s'|s')} \\ \Leftrightarrow \quad & p(s'|s') > p(s'|s), \end{aligned}$$

which, since it needs to hold for all $s, s' \in \{1, \dots, m\}$ with $s' \neq s$, is equivalent to $p(s|s) > p(s|s')$. □

4.5 Robust Bayesian Truth Serum (RBTS)

The Robust Bayesian Truth Serum (RBTS) improves upon the original Bayesian Truth Serum (BTS) in that it is strictly truthful for any number of agents $n \geq 3$. RBTS uses the Shadowing Method from Chapter 3 as a building block. Since the Shadowing Method relies on the mechanism knowing the signal prior, the question is where this signal prior is coming from. As we will see, the signal posterior of a third agent k can play this role.

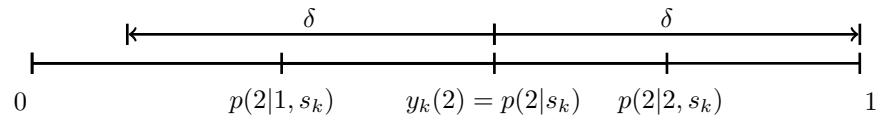


Figure 4.1: An illustration of RBTS with two signals and $S_k = s_k$. Only the beliefs for signal 2 are presented (which imply the beliefs for signal 1). Note that $y_k(2)$ is always strictly in between agent i's two possible second-order posteriors $p(2|1, s_k)$ and $p(2|2, s_k)$.

4.5.1 Mechanism

The *Robust Bayesian Truth Serum (RBTS)* is defined as:

1. Each agent i is asked for two reports:

   - **Signal report**: Let $x_i \in \{1, \ldots, m\}$ be agent i's reported signal.
   - **Prediction report**: Let $y_i \in \mathcal{D}$ be agent i's reported signal posterior.

2. For each agent i, select a *reference* agent $k = i+2$ (modulo n) and calculate

$$p'(\cdot|x_i, y_k) = \begin{pmatrix} y_k(1) - \frac{\delta}{m-1} \\ \cdots \\ y_k(x_i) + \delta \\ \cdots \\ y_k(m) - \frac{\delta}{m-1} \end{pmatrix},$$

   where $\delta = \min\Big(\min_{s\in\{1,\ldots,m\}} y_k(s) \cdot (m-1),\ 1 - \max_{s\in\{1,\ldots,m\}} y_k(s)\Big)$, so that the entries in $p'(\cdot|x_i, y_k)$ are in between 0 and 1.

3. For each agent i, choose *peer* agent $j = i + 1$ (modulo n), and pay agent i:

$$u_i(x_i, x_j, y_k) = \underbrace{R_q\big(p'(\cdot|x_i, y_k), x_j\big)}_{\text{signal score}} + \underbrace{R_q\big(y_i, x_j\big)}_{\text{prediction score}}.$$

   where R_q is the quadratic scoring rule, x_j is the signal report by *peer* agent j, and y_k is the prediction report of *reference* agent k.

4.5.2 Incentive Analysis

The intuition is the following (also compare Figure 4.1): agent i has the same belief model as the other agents. Now imagine that before observing her own signal, the mechanism tells agent i the signal observation $S_k = s_k$ of a third

agent k. Before the actual game starts, agent i would then update her belief about the world states from $\Pr(T)$ to $\Pr(T|S_k = s_k)$ given this signal s_k. At this point, i.e. before she observes her own signal, she then holds the same beliefs as agent k because both agents started with the same belief model and updated their beliefs using the same signal observation. In particular, agent k's signal posterior is then agent i's signal prior.

The mechanism cannot tell agent i the signal of agent k before agent i observes her own signal since one agent would have to start reporting, and that agent would not have a predecessor from whom she could see the signal. Instead, the mechanism ensures that, for any $S_k = s_k$, agent i would want to continue reporting her true signal if the mechanism showed s_k to agent i in hindsight. That is, the effect is the same as if the mechanism could show agent k's signal to agent i, because agent k's signal posterior coincides with agent i's "lifted" signal prior for all possible of agent k's signals. And so, while agent i does not know this lifted signal prior, we can score her with the Shadowing Method using the signal posterior report of agent k as her signal prior.

Analogously to $p(\cdot)$ and $p(\cdot|\cdot)$, let $p(s_j|s_i, s_k) = \Pr(S_j = s_j|S_i = s_i, S_k = s_k)$ be the second-order signal posterior given signals $S_i = s_i$ and $S_k = s_k$. Recall that the identities of agents do not play a role, so that, for example, $p(\cdot|1, 2) = p(\cdot|2, 1)$. Note that the Shadowing Method (Section 3.6) generalizes in the natural way if agent i, before coming to the mechanism, already observed a sequence of other signals. In particular, the Shadowing Method is strictly truthful if agent i already observed signal s'' before coming to the mechanism and if $p(s|s, s'') - p(s, s'') > p(s'|s, s'') - p(s', s'')$ for all $s, s', s'' \in \{1, \dots, m\}$ with $s' \neq s$.

Theorem 4.5. *The Robust Bayesian Truth Serum is strictly BNIC for any $n \geq 3$ if $p(s|s, s'') - p(s|s'') > p(s'|s, s'') - p(s'|s'')$ for all $s, s', s'' \in \{1, \dots, m\}$ with $s' \neq s$.*

Proof. Fix some i, peer j and reference k, and assume agents j and k report truthfully. It needs to be shown that it is the unique best response for agent i to report truthfully. The best response conditions for x_i and y_i can be analyzed for each report type separately, because y_i affects only the prediction score, and x_i affects only the information score. Noting that strict incentives for the prediction report y_i follow directly from the use of the strictly proper quadratic scoring rule (Corollary 3.4), we focus on x_i.

Let $S_k = s_k$ and so $y_k = p(\cdot|s_k)$. Conditioned on agent i's own signal $S_i = s_i$ and this additional signal information, agent i's second-order signal posterior is $p(\cdot|s_i, s_k)$. By Theorem 3.7 it is sufficient that $p(s_i|s_i, s_k) - p(s_i|s_k) > p(s'|s_i, s_k) - p(s'|s_k)$ for all $s_i, s_k, s' \in \{1, \dots, m\}$ with $s' \neq s_i$. □

4.6 The 2-Agent RBTS

In this section, I present an implementation of RBTS that uses only two agents instead of three. Instead of shadowing from the prediction report of a third agent, it shadows from the prediction report of the peer agent while still using the peer agent's signal report as the event that is to be predicted. This double use of the peer agent requires a more complex incentive analysis. In contrast to the regular RBTS from Section 4.5, I only analyze the 2-Agent RBTS for binary signals. A generalization of the analysis to three or more signals is left for future work.

4.6.1 Mechanism

The *2-Agent Robust Bayesian Truth Serum (RBTS)* is defined as:

1. Each agent i is asked for two reports:
   - **Signal report**: Let $x_i \in \{1, 2\}$ be agent i's reported signal.
   - **Prediction report**: Let $y_i \in \mathcal{D}$ be agent i's reported signal posterior.
2. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$, and calculate shadow posterior

$$p'(\cdot|x_i, y_j) = \begin{cases} \begin{pmatrix} y_j(1) + \delta \\ y_j(2) - \delta \end{pmatrix} & \text{if} \quad x_i = 1 \\ \begin{pmatrix} y_j(1) - \delta \\ y_j(2) + \delta \end{pmatrix} & \text{if} \quad x_i = 2, \end{cases}$$

   where $\delta = \min\big(y_j(1), y_j(2)\big)$.
3. For each agent i, pay agent i:

$$u_i(x_i, x_j, y_j) = \underbrace{R_q\big(p'(\cdot|x_i, y_j), x_j\big)}_{\text{signal score}} + \underbrace{R_q\big(y_i, x_j\big)}_{\text{prediction score}} .$$

   where R_q is the quadratic scoring rule, x_j is the signal report by peer agent j, and y_j is the prediction report of peer agent j.

4.6.2 Incentive Analysis

Theorem 4.6. *The 2-Agent RBTS is strictly BNIC if $p(s|s) > p(s|s')$ for all $s, s' \in \{1, 2\}$ with $s' \neq s$.*

Proof. WLOG, let agent i observe signal $S_i = s$. Compare the truthful report $x_i = s$ to the report of some other signal $x_i = s' \neq s$. The particular thing about the analysis of the 2-agent RBTS is that peer agent j is used for both the belief report that is shadowed from as well as for the signal agent i is scored against. When building an expectation, agent i would want to report "all the way" to the vector that reflects the certainty about s_j (in case $S_j = s_j$, her belief about the event $S_j = s_j$ is 1). Let $\mathbf{1}_{s_j}$ denote the vector that has entry 1 in row s_j and 0 in the other row, so that $\mathbf{1}_{s_j}(s) = 1$ if $s_j = s$ and 0 otherwise.

With agent j truthful, we can write $p'(\cdot|\cdot, s_j)$ instead of $p'(\cdot|\cdot, y_j)$ and agent i's shadow posterior (for arbitrary $s, s', s_j \in \{1, 2\}$) is

$$p'(s|s', s_j) = \begin{cases} p(s|s_j) + \delta & \text{if } \; s' = s \\ p(s|s_j) - \delta & \text{else.} \end{cases}$$

Using R_q's quadratic loss, strict truthfulness is thus equivalent to (L_q still takes the report and the belief):

$$\begin{aligned} & \mathbf{E}_{S_j}\big[u_i\big(s, S_j, p(\cdot|S_j)\big)\big] > \mathbf{E}_{S_j}\big[u_i\big(s', S_j, p(\cdot|S_j)\big)\big] \\ \Leftrightarrow \quad & \sum_{s_j=1}^{2} \underbrace{p(s_j|s) \cdot L_q\big(p'(\cdot|s, s_j) \,\big|\, \mathbf{1}_{s_j}\big)}_{\text{Expected loss reporting } x_i = s \text{ if } S_j = s_j, \text{ weighted with probability that } S_j = s_j} \\ < & \sum_{s_j=1}^{2} \underbrace{p(s_j|s) \cdot L_q\big(p'(\cdot|s', s_j) \,\big|\, \mathbf{1}_{s_j}\big)}_{\text{Expected loss reporting } x_i = s' \text{ if } S_j = s_j, \text{ weighted with probability that } S_j = s_j} \end{aligned} \tag{4.1}$$

Note that the loss is between reporting the shadow posterior and reporting $\mathbf{1}_{s_j}$, and not between the shadow posterior and the second-order posterior $p(\cdot|s_i, s_j)$! This is because signal s_j is also the signal agent i is scored against.

Shadow posteriors $p'(s_j|s, s'')$ and $p'(s_j|s', s'')$ differ only in rows s and s', so we can simplify the L_q part:

$$\begin{aligned} & \sum_{s_j=1}^{m} p(s_j|s) \cdot L_q\big(p'(\cdot|s, s_j) \,\big|\, \mathbf{1}_{s_j}\big) < \sum_{s_j=1}^{m} p(s_j|s) \cdot L_q\big(p'(\cdot|s', s_j) \,\big|\, \mathbf{1}_{s_j}\big) \\ \Leftrightarrow \quad & \sum_{s_j=1}^{m} p(s_j|s) \cdot \Big(\big(p'(s|s, s_j) - \mathbf{1}_{s_j}(s)\big)^2 + \big(p'(s'|s, s_j) - \mathbf{1}_{s_j}(s')\big)^2\Big) \\ < & \sum_{s_j=1}^{m} p(s_j|s) \cdot \Big(\big(p'(s|s', s_j) - \mathbf{1}_{s_j}(s)\big)^2 + \big(p'(s'|s', s_j) - \mathbf{1}_{s_j}(s')\big)^2\Big) \end{aligned} \tag{4.2}$$

To break down $\mathbf{1}_{s_j}$ and simplify this equation with regard to the summation, I distinguish two cases:

1. $S_j = s$, which occurs with probability $p(s|s)$.

Reporting $x_i = s$ (left side of inequality):

$$\begin{aligned}&p(s|s) \cdot \Big(\big(p'(s|s,s) - \mathbf{1}_s(s)\big)^2 + \big(p'(s'|s,s) - \mathbf{1}_s(s')\big)^2 \Big) \\ =&p(s|s) \cdot \Big(\big(1 - p'(s|s,s)\big)^2 + p'(s'|s,s)^2 \Big)\end{aligned}$$

Reporting $x_i = s'$ (right side of inequality):

$$\begin{aligned}&p(s|s) \cdot \Big(\big(p'(s|s',s) - \mathbf{1}_s(s)\big)^2 + \big(p'(s'|s',s) - \mathbf{1}_s(s')\big)^2 \Big) \\ =&p(s|s) \cdot \Big(\big(1 - p'(s|s',s)\big)^2 + p'(s'|s',s)^2 \Big)\end{aligned}$$

2. $S_j = s'$, which occurs with probability $p(s'|s) = 1 - p(s|s)$:

Reporting $x_i = s$:

$$\begin{aligned}&p(s'|s) \cdot \Big(\big(p'(s|s,s') - \mathbf{1}_{s'}(s)\big)^2 + \big(p'(s'|s,s') - \mathbf{1}_{s'}(s')\big)^2 \Big) \\ =&p(s'|s) \cdot \Big(p'(s|s,s')^2 + \big(1 - p'(s'|s,s')\big)^2 \Big)\end{aligned}$$

Reporting $x_i = s'$:

$$\begin{aligned}&p(s'|s) \cdot \Big(\big(p'(s|s',s') - \mathbf{1}_{s'}(s)\big)^2 + \big(p'(s'|s',s') - \mathbf{1}_{s'}(s')\big)^2 \Big) \\ =&p(s'|s) \cdot \Big(p'(s|s',s')^2 + \big(1 - p'(s'|s',s')\big)^2 \Big)\end{aligned}$$

Putting this together, we obtain (with $s' \neq s$):

$$\begin{aligned}&\sum_{s_j=1}^{2} p(s_j|s) \cdot L_q\big(p'(\cdot|s) \,\big|\, \mathbf{1}_{s_j}\big) < \sum_{s_j=1}^{2} p(s_j|s) \cdot L_q\big(p'(\cdot|s') \,\big|\, \mathbf{1}_{s_j}\big) \\ \Leftrightarrow \quad & p(s|s) \cdot \Big(\big(1 - p'(s|s,s)\big)^2 + p'(s'|s,s)^2 \Big) \\ & + \big(1 - p(s|s)\big) \cdot \Big(p'(s|s,s')^2 + \big(1 - p'(s'|s,s')\big)^2 \Big) \\ < \quad & p(s|s) \cdot \Big(\big(1 - p'(s|s',s)\big)^2 + p'(s'|s',s)^2 \Big) \\ & + \big(1 - p(s|s)\big) \cdot \Big(p'(s|s',s')^2 + \big(1 - p'(s'|s',s')\big)^2 \Big)\end{aligned}$$

$$
\begin{aligned}
\Leftrightarrow \quad & p(s|s) \cdot \left(\Big(1 - \big(p(s|s) + \delta\big)\Big)^2 + \big(p(s'|s) - \delta\big)^2 \right) \\
& + \big(1 - p(s|s)\big) \cdot \left(\big(p(s|s') + \delta\big)^2 + \Big(1 - \big(p(s'|s') - \delta\big)\Big)^2 \right) \\
< \quad & p(s|s) \cdot \left(\Big(1 - \big(p(s|s) - \delta\big)\Big)^2 + \big(p(s'|s) + \delta\big)^2 \right) \\
& + \big(1 - p(s|s)\big) \cdot \left(\big(p(s|s') - \delta\big)^2 + \Big(1 - \big(p(s'|s') + \delta\big)\Big)^2 \right) \\
\underbrace{\Leftrightarrow}_{\delta > 0} \quad & p(s|s) > p(s|s')
\end{aligned}
$$

□

4.7 Comparison of Mechanisms

Analogous to the 1/prior mechanism and the Shadowing Method in Chapter 3, the 1/posterior BTS and the Robust Bayesian Truth Serum (RBTS) require different conditions for strict truthfulness (compare Theorem 4.4 and Theorem 4.5). Therefore, as in Section 3.7, the first question is how these conditions compare to each other. As we will see, the conditions for strict truthfulness are identical for binary signals and incomparable for more than two signals. The truthfulness conditions are not the only property where the mechanisms differ. In particular, depending on which version of RBTS is used, it requires 2 or 3 agents, whereas the 1/posterior BTS always requires only 2 agents. A perhaps more important distinction is that the ex post payments of RBTS are bounded whereas the ex post payments computed by the 1/posterior BTS can be arbitrarily large.

4.7.1 Truthfulness Conditions

Theorem 4.7. *With binary signals, the conditions for strict truthfulness required by the 1/posterior BTS and the 2-Agent Robust Bayesian Truth Serum are identical.*

Proof. The statement follows directly from Theorems 4.4 and 4.6. □

Theorem 4.8. *With more than two signals, the conditions for strict truthfulness required by the 1/posterior BTS and the Robust Bayesian Truth Serum (RBTS) are incomparable.*

Proof. The proof is by example. An example of beliefs, where RBTS is strictly

truthful and the 1/posterior BTS is not, is the following with $m = 3$ signals:

$$\Pr(T) = \begin{pmatrix} 0.1 \\ 0.8 \\ 0.1 \end{pmatrix}$$
$$\Pr(S|T) = \begin{pmatrix} 0.5 & 0.2 & 0.35 \\ 0.3 & 0.75 & 0.25 \\ 0.2 & 0.05 & 0.4 \end{pmatrix}.$$

This results in signal prior

$$p(\cdot) = \Pr(S|T) \times \Pr(T) = \begin{pmatrix} 0.245 \\ 0.655 \\ 0.1 \end{pmatrix}$$

and signal posterior matrix

$$p(\cdot|\cdot) = \begin{pmatrix} 0.28265306 & 0.21946565 & 0.32 \\ 0.58673469 & 0.71030534 & 0.46 \\ 0.13061224 & 0.07022901 & 0.22 \end{pmatrix}.$$

The second-order signal posterior matrices are

$$p(\cdot|\cdot, S_k = 1) = \begin{pmatrix} 0.33483755 & 0.24043478 & 0.359375 \\ 0.49909747 & 0.6726087 & 0.390625 \\ 0.16606498 & 0.08695652 & 0.25 \end{pmatrix}$$
$$p(\cdot|\cdot, S_k = 2) = \begin{pmatrix} 0.24043478 & 0.20781838 & 0.27173913 \\ 0.6726087 & 0.73457818 & 0.5826087 \\ 0.08695652 & 0.05760344 & 0.14565217 \end{pmatrix}$$
$$p(\cdot|\cdot, S_k = 3) = \begin{pmatrix} 0.359375 & 0.27173913 & 0.36363636 \\ 0.390625 & 0.5826087 & 0.30454545 \\ 0.25 & 0.14565217 & 0.33181818 \end{pmatrix}.$$

First note that the signal posterior matrix does not satisfy $p(s|s) > p(s|s')$ for all $s, s' \in \{1, \ldots, m\}$ with $s' \neq s$ since $p(1|1) = 0.28265306 \not> 0.32 = p(1|3)$. Comparing the signal posterior matrix with the second-order signal posterior matrix, one can verify that $p(s|s, s'') - p(s|s'') > p(s'|s, s'') - p(s'|s'')$ for all $s, s', s'' \in \{1, \ldots, m\}$ with $s' \neq s$.

Similarly, an example of beliefs, where the 1/posterior BTS is strictly truth-

ful but RBTS is not is the following with $m = 3$ signals:

$$\Pr(T) = \begin{pmatrix} 0.55 \\ 0.2 \\ 0.25 \end{pmatrix}$$

$$\Pr(S|T) = \begin{pmatrix} 0.45 & 0.75 & 0.1 \\ 0.35 & 0.2 & 0.75 \\ 0.2 & 0.05 & 0.15 \end{pmatrix}.$$

This results in signal prior

$$p = \begin{pmatrix} 0.4225 \\ 0.42 \\ 0.1575 \end{pmatrix}$$

and signal posterior matrix

$$p(\cdot|\cdot) = \begin{pmatrix} 0.53579882 & 0.32232143 & 0.38571429 \\ 0.3204142 & 0.51428571 & 0.43571429 \\ 0.14378698 & 0.16339286 & 0.17857143 \end{pmatrix}$$

which satisfies $p(s|s) > p(s|s')$.

The second-order signal posterior matrices are:

$$p(\cdot|\cdot, S_k = 1) = \begin{pmatrix} 0.59522363 & 0.46800554 & 0.4654321 \\ 0.279873 & 0.37216066 & 0.35617284 \\ 0.12490337 & 0.1598338 & 0.17839506 \end{pmatrix}$$

$$p(\cdot|\cdot, S_k = 2) = \begin{pmatrix} 0.46800554 & 0.23324653 & 0.31530055 \\ 0.37216066 & 0.60486111 & 0.50956284 \\ 0.1598338 & 0.16189236 & 0.17513661 \end{pmatrix}$$

$$p(\cdot|\cdot, S_k = 3) = \begin{pmatrix} 0.4654321 & 0.31530055 & 0.38533333 \\ 0.35617284 & 0.50956284 & 0.42733333 \\ 0.17839506 & 0.17513661 & 0.18733333 \end{pmatrix}.$$

This model does not satisfy the RBTS condition $p(s|s, s'') - p(s|s'') > p(s'|s, s'') - p(s'|s'')$ for all $s, s', s'' \in \{1, \ldots, m\}$ with $s' \neq s$ because $0.0346080794799 = 0.178395061728 - 0.143786982249 = p(3|3, 1) - p(3|1) <= p(2|3, 1) - p(2|1) = 0.356172839506 - 0.320414201183 = 0.0357586383227$. □

4.7.2 Other Properties

Theorem 4.8 states that the conditions required for strict truthfulness of the 1/posterior BTS and RBTS are incomparable. However, the truthfulness conditions are not the only difference between the two mechanisms. For example, the number of required agents is $n \geq 3$ for multi-signal RBTS and only $n \geq 2$ for multi-signal 1/posterior BTS (both require $n \geq 2$ for binary signals). This is not a major difference since one would expect the designer to ask for three or more reports for information aggregation purposes anyways in these complex settings, where minority opinions may be correct.

A more important difference between the two mechanisms stems from the fact that—in contrast to the mechanisms of Chapter 3, where the known signal prior is used to score agent i—the beliefs y_j (in 1/posterior BTS) or y_k (in RBTS) are reported by agents. This has implications on the mechanisms' payments in the worst case. In particular, Theorems 4.9 and 4.10 show that while the ex post payments using 1/posterior BTS are unbounded, i.e. can get arbitrarily large, the ex post payments using RBTS are bounded. Moreover, in RBTS, this upper bound can be set to any value chosen by the designer. Bounded ex post payments are crucial in practice because the designer can set an upper bound on its willingness to pay. For example, the designer may want to cap payments for a single report at \$0.50. With 1/posterior BTS, there is no such cap and the designer could end up having to pay any high amount (it is unbounded, so the worst case is an infinite amount). Furthermore, unbounded ex post payments may also increase a mechanism's susceptibility to collusion since the agents' profit from coordinating reports would be infinite.

Theorem 4.9. *The ex post scores in the 1/posterior BTS can be unbounded.*

Proof. Imagine there are $m = 2$ signals and agents i and j report:

$$x_i = x_j = 1$$
$$y_i = y_j = \begin{pmatrix} \varepsilon \\ 1 - \varepsilon \end{pmatrix}$$

Then, the ex post score to each agent is $\frac{1}{\varepsilon}$ which goes to infinity for $\varepsilon \to 0$. □

Theorem 4.10. *The ex post scores in the Robust Bayesian Truth Serum are in* $[0, 2]$ *for any reports from agents including* $y_i(s) = 0$ *for any* $s \in \{1, \ldots, m\}$*, and thus RBTS is ex post individually rational and numerically robust.*

Proof. The quadratic scoring rule $R_q(y, \omega)$ is well-defined for any input $y \in \mathcal{D}$ and $\omega \in \{0, \ldots, m\}$, and generates scores on $[0, 1]$. The inputs to R_q for computing the information score are $y := p'(\cdot|x_i, y_k) \in \mathcal{D}$ and $\omega := x_j \in \{0, \ldots, m\}$.

The inputs for computing the prediction score are $y := y_i \in \mathcal{D}$ and $\omega := x_j \in \{0, \ldots, m\}$. □

This has a nice implication: for a designer with a budget $B > 0$, a straightforward extension of RBTS is to multiply R_q with a positive scalar $\alpha > 0$ to implement a mechanism that conforms with any budget constraint, since the total ex post cost is upper-bounded by $2\alpha n$.

A simple randomized extension of multi-signal RBTS achieves constant *ex post* budget of $B > 0$ for groups of $n \geq 4$ by randomly excluding an agent from the population, running RBTS with budget $B > 0$ on the remaining $n - 1$ agents, and redistributing whatever remains from B to the excluded agent. This extension to RBTS remains strictly incentive compatible when the agents do not know which of them is the excluded agent. While multiple equilibria cannot be avoided in peer prediction settings without trusted reports [Jurca and Faltings, 2005; Waggoner and Chen, 2013], this randomized extension ensures that the agents' scores in the truthful equilibrium cannot be less than in any other equilibrium. Moreover, by sacrificing *ex post* individual rationality, the same technique can be used to implement a mechanism with $B = 0$.

In contrast to BTS, RBTS easily adapts to *online* polling settings, where the mechanism publishes partial information about reports as agents arrive. Since RBTS achieves incentive compatibility for any group with $n \geq 3$ agents, the mechanism can sequentially score groups of three, and subsequently release their reports. The 1/posterior BTS also has this property.

4.8 Conclusion

In this chapter, I introduced Bayesian Truth Serum mechanisms that take the same inputs as the original Bayesian Truth Serum but achieve strict Bayes-Nash incentive compatibility for small populations. My Robust Bayesian Truth Serum (RBTS) is strictly truthful for every number of agents $n \geq 3$, the 2-Agent RBTS is strictly truthful for every number of agents $n \geq 2$, but restricted to binary signals.

I believe RBTS can have practical impact, providing a more principled approach to incentivize small groups of workers on crowdsourcing platforms such as Amazon Mechanical Turk, where the original Bayesian Truth Serum has already been shown to be useful for quality control [Shaw et al., 2011].

There are two interesting directions for future work. The first is to generalize the analysis of the 2-Agent RBTS to more than two signals. The second is to extend the analysis to different agent types, reflecting how different parts of the population evaluate the same experience. For example, elderly couples

and college students may have different understandings of what constitutes a good hotel. In fact, booking.com[2] already recognizes this and lets users sort hotel reviews according to different reviewer types, such as "mature couples" and "solo travelers." When adapting RBTS to this model, the format of the signal and the prediction report could stay the same. The prediction report could also continue to be scored as is. What would need to change is the incentive analysis of the signal report, incorporating the distribution over reviewer types and their respective signal posteriors. It will be interesting to see under which conditions the current RBTS design continues to be truthful in this model. Simpler approaches that penalize disagreement of the prediction reports amongst agents who report the same signal do not seem amenable to these heterogeneous settings.

[2] `http://www.booking.com`

Chapter 5

Subjective-Prior Peer Prediction

The classical peer prediction method (Chapter 2) is strictly truthful but relies on a common belief model, shared by all agents and the mechanism. The Shadowing Method (Chapter 3) relaxes this assumption, in that the agents share a common belief model but the mechanism only needs to know the signal prior. Bayesian Truth Serum mechanisms (Chapter 4) further relax the common knowledge required by the mechanism in that the commonly-held belief model need not be known by the mechanism. However, all mechanisms we have seen so far still assume that the belief model is shared among all agents. For example, while Bayesian Truth Serum mechanisms do not require the mechanism to know the belief model, all agents still need to share the same model.

To support these assumptions, the authors of the classical peer prediction method (Chapter 2) suggest that, in the context of feedback about a product or service, the rating history can be leveraged in order to allow the mechanism to estimate the belief model. This, however, leaves open the question as to how the rating history itself was built: either people reported truthfully without an incentive-compatible mechanism in place (and there is no design problem) or people reported dishonestly, in which case the mechanism cannot use these reports to learn the correct belief model. Beyond this difficulty, there remains the concern that every user must share the same belief model, something that seems unreasonable in practice.

Relaxing this assumption of a common belief model is the focus of this chapter, and I introduce two mechanisms that provide strict incentives for truthful reports in equilibrium while allowing participants to have subjective and private models. Both mechanisms ask a user for two reports: one before she observes

her signal and one afterwards. The ability to enforce this *temporal separation* is critical but seems reasonable in many applications. For example, a travel site could ask a user for her opinion about a hotel at the time of booking and then again after her stay.

I first introduce the *basic subjective-prior peer-prediction mechanism (BSPP)*, that requires an agent to report two belief reports about the signal that another agent will receive, one before and one after receiving her own signal. BSPP is strictly incentive compatible, and infers the agent's signal from the change in her belief reports. Building on this, I introduce and analyze the equilibrium of a mechanism in which an agent's first belief report is followed by only a signal report. In this candidate mechanism, truthful reporting of the signal, but not of the belief, is an equilibrium. Computing an agent's optimal belief misreport, I then construct the strictly incentive compatible *shadow subjective-prior peer-prediction mechanism (SSPP)* via an application of the revelation principle [Gibbard, 1973; Green and Laffont, 1977; Myerson, 1979, 1981], simulating this misreport on behalf of an agent. Moreover, I present a special case of SSPP that has a very simple and intuitive form.

The main technical innovation is to apply the Shadowing Method (Chapter 3) to an agent's *own prediction report*, which requires a more complex analysis because it introduces sequential reasoning on behalf of the agent. More specifically, since the signal report of an agent will be scored using the Shadowing Method applied to her own prediction report, when reporting her prediction report, she not only considers her prediction score but already incorporates the implications on her signal score.

The only knowledge SSPP requires about the agents' belief models is that the effect an agent's signal has on her signal posterior is bounded away from zero. This technical requirement in regard to the minimal informativeness of signals given the agents' beliefs is not required for the strict incentives of BSPP, in which the second report is a belief report rather than a signal report.

In moving from a common knowledge setting to one with private and subjective belief models, an important consideration is the amount of additional information that must be elicited from a participant over and above a signal report. Indeed, in the original paper of the classical peer prediction method, Miller et al. [2005] had suggested the possibility of incentive compatible peer prediction with subjective and private belief models. In a brief treatment, they proposed an approach in which, in addition to her own signal, a user also reports her belief on the world state and her belief on the probability of receiving each possible signal, conditioned on each possible state. In comparison, both BSPP and SSPP are considerably simpler with respect to the reporting costs, and thus likely more practical. In fact, my analysis suggests a trade-off between

the robustness of incentive properties and the reporting requirements, given that SSPP but not BSPP requires the technical requirement on the minimal informativeness of signals.

A limitation of both BSPP and SSPP relative to the original peer prediction method is that in its current form, its application is restricted to domains with binary signals. However, many interesting applications of peer prediction mechanisms are to settings with binary signals. For example, blogs and online forums allow users to vote whether a post was helpful or not. Similarly, social networking websites, such as Facebook and Google+, allow users to "like" or "+1" other users' comments. Hotel booking websites, such as Expedia and Hotwire, ask customers after their stay whether they "would recommend this hotel to a friend," and the report as to whether a given website contains offensive content, is binary, too. I leave the extension to multiple signals for future work.

The remainder of this chapter is organized as follows. After referencing related work specific to the subjective model used in this chapter in Section 5.1, I explain the difference of the model used in this chapter as compared to the standard model in Section 5.2. In Section 5.3, I then introduce and analyze the basic subjective-prior peer prediction mechanism (BSPP), which asks for two belief reports. In Section 5.4, I introduce the candidate shadow subjective-prior peer prediction mechanism (Candidate SSPP), a first attempt at a subjective-prior mechanism that requires only a belief and a signal report. While Compact SSPP's belief report is not truthful, its incentive analysis leads to the design of the shadow subjective-prior peer prediction mechanism (SSPP) that I introduce in Section 5.5. An observation in SSPP's payment rule leads to its compact, intuitive form (Compact SSPP), which I describe in Section 5.5.2. I conclude the chapter with a discussion of considerations when applying these mechanisms and interesting directions for future work in Section 5.6.

5.1 Related Work

In addition to the related work on peer prediction mechanisms introduced in Chapters 2 to 4, there is additional related work particular to this chapter.

Various notions of subjective, self-confirming and conjectural equilibria appear in the game theory literature, although normally studied in repeated contexts [e.g. Kalai and Lehrer, 1993; Battigali et al., 1992]. The most related concept to the solution concept used in this chapter is that of Rubinstein and Wolinsky [1984], who propose *rationalizable conjectural equilibria* in the context of a one-shot game, but without my notion of *ex post* robustness and strict uncertainty about the other agents' belief types. In addition, they require agent

observations in regard to the play of others and consistency of beliefs relative to these observations. I model a single interaction between an agent and a peer prediction mechanism, and the agent is unable to make any observations about the actions of her peer agent (against which she is scored.)

5.2 Model

The model of this chapter is the binary-signal variation of the standard model (Section 2.1), where each agent has her own subjective belief model, which is unknown to the mechanism.

Definition 12. The subjective belief model of agent i is referred to as agent i's *belief type* $\theta_i \in \Theta$, and denoted with a subscript indicating the agent, i.e. $\Pr_i(T)$ and $\Pr_i(S|T)$.

Analogously, the shorthand notation for subjective signal prior and subjective signal posteriors resulting from the agent's belief type also takes a subscript indicating the agent, i.e. $p_i(\cdot)$ and $p_i(\cdot|\cdot)$. In this chapter, I consider only binary signals, i.e. $m = 2$, and I will refer to signal 1 as the *low* signal and to signal 2 as the *high* signal. Note that with binary signals, $p_i(2)$ and $p_i(2|\cdot)$ fully capture agent i's subjective beliefs about her peer agent's signal, because $p_i(1) = 1 - p_i(2)$ and $p_i(1|s) = 1 - p_i(2|s)$ for any $s \in \{1, 2\}$.

5.3 Basic Subjective-Prior Peer Prediction Mechanism (BSPP)

In handling private and subjective belief models, in place of the usual Bayes-Nash equilibrium analysis of peer prediction mechanisms, I analyze the subjective equilibrium of a mechanism. Informally, this requires that each agent best-responds to the strategy of every other agent given her own subjective belief model, and given strict uncertainty about the belief models of other agents. Put differently, agents form subjective beliefs about the signals received by other agents but are not required to have beliefs about the subjective beliefs of other agents.

There are two problems in extending the classical peer prediction method to incorporate subjective belief models. First, it is no longer sufficient for the mechanism to only ask for signal reports because it would not be able to infer signal posteriors. This could be solved by eliciting signal posteriors, but then we would run into a second problem: without the belief model being common knowledge, eliciting only the signal posterior does not enable the mechanism to

infer the agent's signal. In addition to the signal being the only objective piece of information, eliciting the signal is crucial because it is used as the event that another agent shall predict.

Example 6. *Consider the numbers from Example 1 in Section 2.2, i.e.* $m = 2$, $\Pr(T = 2) = 0.3$, $\Pr(S = 2 \mid T = 2) = 0.6$, *and* $\Pr(S = 2 \mid T = 1) = 0.1$. *Assume the mechanism asked for signal posteriors and agent* i *observed* $S_i = 2$. *Even if she truthfully reported*

$$y_i = \begin{pmatrix} p(1|2) \\ p(2|2) \end{pmatrix} = \begin{pmatrix} 0.54 \\ 0.46 \end{pmatrix},$$

the mechanism could not infer agent i*'s signal without knowledge of her subjective belief model because the same signal posterior could also stem from a low signal, i.e.* $S_i = 1$, *in a setting with* $\Pr(T = 2) = 0.853$ *and the same conditional signal beliefs.*

As a possible solution for settings with private and subjective belief models, the authors of the classical peer prediction method (Chapter 2) briefly discuss the possibility of a direct-revelation approach where the agents are asked to report both the entire belief model ($\Pr_i(T)$ and $\Pr_i(S|T)$) and the signal itself. In fact, this approach is not strictly truthful if all information is reported simultaneously.

Example 7. *Consider again the numbers from Example 1 in Section 2.2: if agent* i *observes a high signal, i.e.* $S_i = 2$, *a truthful report of* $\Pr(T = 2) = 0.3$, $\Pr(S = 2 \mid T = 2) = 0.6$, *and* $\Pr(S = 2 \mid T = 1) = 0.1$, *in addition to the truthful signal report* $x_i = 2$, *yields the same payment as a misreport of a low signal, i.e.* $x_i = 1$, *with beliefs* $\Pr(T = 2) = 0.853$, $\Pr(S = 2 \mid T = 2) = 0.6$ *and* $\Pr(S = 2 \mid T = 1) = 0.1$, *because the induced signal posteriors of these models are the same (compare Example 6).*

Of course, if strict truthfulness is not required, the naive mechanism that asks only for a signal report and pays each agent a constant amount independent of the reported signal is a much simpler, weakly truthful solution.

While the authors of the classical peer prediction method do not mention this, their direct-revelation approach can be made strictly truthful by temporal separation. The mechanism must ensure that the agent reports her subjective probabilistic model *before* receiving her signal. Nevertheless, this direct approach appears impractical because of its high reporting costs. Observe that in the case of two states and two signals, an agent has to report three probabilities and a signal. These reporting costs grow with the number of states, so that with three states and two signals, it would already require each agent to report five

	Belief Model	
	common	subjective
public	Classical Peer Prediction	(Classical Peer Prediction)
private	Bayesian Truth Serums	this chapter

Table 5.1: Overview of belief models assumed by different peer prediction mechanisms.

probabilities and a signal. Therefore, a different approach is required for settings with subjective belief models, in order to design a mechanism that is both strictly incentive compatible, and feasible with respect to the agents' reporting costs.

I deviate from the classical peer prediction method in two aspects. First, every agent has her own subjective belief type in regard to the world state and the way in which signals are generated given each state. Second, this belief type is private to an agent and not known by other agents or the mechanism. The difficulty comes from the second relaxation: if the mechanism knew each agent's subjective beliefs, it could still compute the possible posterior beliefs for the other agent's signal and the classical peer prediction method could be applied.

Bayesian Truth Serum mechanisms (Chapter 4) provide a solution when all agents share a common belief model, but where this model is not known by the mechanism. In addition to the signal report, these mechanisms ask agents to report their posterior signal beliefs and score agents on the basis of both of these reports. Table 5.1 provides a summary of the different settings. In the remainder of this thesis, I will use "subjective" to mean private and subjective.

In what follows I provide a first proposal, the *basic subjective-prior peer-prediction mechanism (BSPP)*, for the setting of private and subjective belief models.

5.3.1 Mechanism

The *basic subjective-prior mechanism (BSPP)* is defined as:

1. Ask agent i for her signal prior report $y_i \in \mathcal{D}$.
2. Agent i observes signal $S_i = s_i$.
3. Ask agent i for her signal posterior report $z_i \in \mathcal{D}$, enforcing $z_i \neq y_i$.
4. Infer agent i's implicit signal report x_i by applying

$$x_i = \begin{cases} 2, & \text{if} \quad z_i(2) > y_i(2) \\ 1, & \text{if} \quad z_i(1) > y_i(1) \end{cases}$$

5. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$ and pay agent i:

$$u_i\big(y_i, z_i, x_j\big) = R\big(y_i, x_j\big) + R\big(z_i, x_j\big),$$

where R is a strictly proper scoring rule, and x_j is the inferred signal report of peer agent j.

As we will see, the true signal prior and the true signal posterior cannot be the same given stochastic relevance, so that $z_i \neq y_i$ is not restrictive for a truthful agent.[1]

5.3.2 Incentive Analysis

The equilibrium concept used in this chapter does not require agent i to form beliefs about the belief types of other agents.

Definition 13. Agent i's *strategy* σ_i in the BSPP mechanism consists of two components. Component $\sigma_i(\theta_i)$ takes one argument with $\sigma_i : \Theta \to [0, 1]$, defining agent i's first report for every possible belief type, and component $\sigma_i(\theta_i, s_i)$ takes two arguments with $\sigma_i : \Theta \times \{1, 2\} \to [0, 1]$, defining agent i's report for every possible belief type and every possible signal observation $S_i = s_i$.

Definition 14. Strategy profile $(\sigma_1^*, \ldots, \sigma_n^*)$ is an *ex post subjective equilibrium* of the BSPP mechanism with n agents if

$$\begin{aligned}
&\mathbf{E}_{S_i,S_j}\Big[u_i\big(\sigma_i^*(\theta_i), \sigma_i^*(\theta_i, S_i), S_j\big)\Big] + \mathbf{E}_{S_j}\Big[u_i\big(\sigma_i^*(\theta_i), \sigma_i^*(\theta_i, s_i), S_j\big) \mid S_i = s_i\Big] \\
\geq\ &\mathbf{E}_{S_i,S_j}\Big[u_i\big(\sigma_i(\theta_i), \sigma_i(\theta_i, S_i), S_j\big)\Big] + \mathbf{E}_{S_j}\Big[u_i\big(\sigma_i(\theta_i), \sigma_i(\theta_i, s_i), S_j\big) \mid S_i = s_i\Big]
\end{aligned}$$

for all $i \in \{1, \ldots, n\}$, $j = i + 1 \,(\text{modulo } n)$, all $s_i \in \{1, 2\}$, and all $\sigma_i \neq \sigma_i^*$. It is a strict ex post subjective equilibrium if the inequality is strict.

In this equilibrium concept, each agent i is best-responding to the strategy of every other agent given knowledge of her own type (i.e., her own subjective belief model) and given common knowledge of rationality. The equilibrium is *subjective* because it allows for each agent to have a distinct belief type, and *ex post* because it allows for strict uncertainty in regard to the types of other agents. Ex post subjective equilibrium is strictly more general than Bayes-Nash equilibrium (BNE) because it coincides with BNE when all agents have the

[1] To keep the user interface simple, a practical deployment might allow the two reports from an agent to be equal and still pay the agent as described. In this case, the method would skip agent i's role as peer agent for agent $i - 1 \,(\text{modulo } n)$. In the extreme case, where there is some agent i for which all other agents $j \neq i$ report $z_j = z_j$, a practical deployment could score i against a random signal. It bears emphasis that these details do not affect the equilibrium analysis, but are all robustness issues in regard to a practical deployment.

same belief type. Since agent i's best response still depends on peer agent j's report, ex post subjective equilibrium is less general than dominant strategy implementation. The solution concept used in this chapter is thus strictly in between these two well-known solution concepts.

We are interested in *ex post subjective incentive compatible* mechanisms, where all agents i play the truthful strategy $\sigma_i^1(\theta_i) = p_i(2)$ and $\sigma_i^2(\theta_i, s_i) = p_i(2|s_i)$ is an equilibrium. That is, every agent reports her true signal prior $p_i(2) = \Pr_i(S_i = 2)$ and then her true signal posterior $p_i(2|s_i) = \Pr_i(S_j = 2 \mid S_i = s_i)$. A mechanism is *strictly* incentive compatible when the equilibrium is strict.

It is critical for the incentive properties of BSPP that Step 1 happens before the agent observes signal S_i and Step 3 happens after the agent has observed signal S_i. This is the property of temporal separation. In establishing incentive compatibility, we need Lemma 5.1.

Lemma 5.1. *Stochastic relevance of θ_i implies $p_i(2|2) > p_i(2)$ and $p_i(2|1) < p_i(2)$.*

Proof. The statement follows directly from Corollary 3.6 and Theorem 3.9. □

In words, agent i's belief that another agent receives signal 2 strictly increases from her signal prior in the event that she observes signal 2. Analogously, her belief that another agent receives signal 2 strictly decreases relative to her signal prior if she observes signal 1.

Theorem 5.2. *BSPP is strictly ex post subjective incentive compatible if stochastic relevance holds for every agent's belief type and given temporal separation.*

Proof. Assume peer agent j is truthful. First, for agent j's inferred signal report it holds that $x_j = s_j$ with $S_j = s_j$. To see this, verify that by Lemma 5.1 it holds that $z_j(2) = p_j(2|s_j) > p_j(2) = y_j(2)$ if and only if $s_j = 2$, and that $z_j(2) = p_j(2|s_j) < p_j(2) = y_j(2)$ if and only if $s_j = 1$. It follows that agent i has strict incentives to report y_i and z_i truthfully, with respect to her subjective belief type because her score is the sum of two proper scoring rules applied to x_j and recognizing that the inference in regard to x_i does not affect agent i's score. □

5.3.3 Individual Rationality

In contrast to the classical peer prediction method (Chapter 2), the particular choice of scoring rule matters as to whether BSPP provides individual rationality. In particular, the logarithmic scoring rule $R_{\log}$ (Section 2.5.1) cannot

provide individual rationality for BSPP. This is because an agent's signal posterior $p_i(2|s_i)$ can take values arbitrarily close to 0, and thus there is no suitable constant the mechanism could add to always make the score non-negative.

Instead, we can adopt a strictly proper scoring rule that guarantees non-negative values for every possible report, so that there is no need for adding a lower bound. An example for such a rule is the quadratic scoring rule R_q, which was introduced in Section 3.4. For binary events, i.e. $\Omega = \{1, 2\}$, the normalized quadratic scoring rule simplifies to:

$$\begin{aligned} R_q(y, \omega) &= y(\omega) - 0.5\Big(y(\omega)^2 + \big(1 - y(\omega)\big)^2\Big) + 0.5 \\ &= 2y(\omega) - y(\omega)^2, \end{aligned} \tag{5.1}$$

which, written out for $\omega = 1, 2$ and with the agent only reporting her belief for a high signal $y(2)$, results in:

$$\begin{aligned} R_q\big(y(2), \omega = 2\big) &= 2y(2) - y(2)^2 \\ R_q\big(y(2), \omega = 1\big) &= 1 - y(2)^2. \end{aligned} \tag{5.2}$$

Example 8. *Consider again the ad exchange example from Section 2.2 with numbers* $m = 2$, $\Pr_i(T = 2) = 0.3$, $\Pr_i(S = 2 \mid T = 2) = 0.6$, *and* $\Pr_i(S = 2 \mid T = 1) = 0.1$. *The procedure using BSPP together with the quadratic scoring rule* R_q *is then as follows:*

1. *Worker* i *accepts the task that was posted on Amazon Mechanical Turk by the ad exchange. She is asked for her signal prior report of observing violence, and she truthfully reports*

$$y_i(2) = 0.25.$$

2. *A website that may or may not contain violent content is shown to worker* i *and, after looking at it carefully, she is asked to report her signal posterior of some other agent observing violence on that same website. She did not observe violence, i.e.* $S_i = 1$, *and so she updates her signal posterior to* $p_i(2|1) = 0.18$ *and truthfully reports*

$$z_i(2) = 0.18.$$

3. *Another worker* $j \neq i$ *also follows Steps 1 and 2, with potentially different beliefs and experiences.*

4. *Worker* i *is scored against worker* j*'s implicitly reported signal* x_j. *Assum-*

ing agent j also did not observe violent content on the website and reported both signal prior and signal posterior truthfully, so that the implicit signal report is $x_j = 1$, agent i is paid

$$R_q(0.25, 1) + R_q(0.18, 1) = 2 - 0.25^2 - 0.18^2 = 1.905.$$

5.4 Candidate Shadow Subjective-Prior Mechanism (Candidate SSPP)

In this section, I modify the BSPP mechanism so that an agent's second report is a signal report rather than a belief report. I first introduce a candidate shadow peer-prediction mechanism, in which the equilibrium provides strict incentives for truthful signal reports but not for truthful belief reports. Section 5.4.2 provides an analysis of this mechanism, identifying the optimal deviation from the true prior belief report. By making an appeal to the *revelation principle* from mechanism design theory [Gibbard, 1973; Green and Laffont, 1977; Myerson, 1979, 1981], I then construct the shadow subjective-prior peer-prediction mechanism (Section 5.5), which is strictly incentive compatible for both reports.

5.4.1 Mechanism

The *candidate shadow subjective-prior mechanism (Candidate SSPP)* is defined as:

1. (Stage 1) Ask agent i for her signal prior report $y_i \in \mathcal{D}$.

2. Agent i observes signal $S_i = s_i$.

3. (Stage 2) Ask agent i for her signal report $x_i \in \{1, 2\}$, and calculate shadow posterior

$$p'(\cdot|x_i, y_i) = \begin{cases} \begin{pmatrix} y_i(1) + \delta \\ y_i(2) - \delta \end{pmatrix} & \text{if} \quad x_i = 1 \\ \begin{pmatrix} y_i(1) - \delta \\ y_i(2) + \delta \end{pmatrix} & \text{if} \quad x_i = 2, \end{cases}$$

   where $\delta > 0$ is a parameter of the mechanism.

4. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$ and pay agent i:

$$u_i\big(y_i, x_i, x_j\big) = R_q\big(y_i, x_j\big) + R_q\Big(p'\big(\cdot|x_i, y_i\big), x_j\Big)$$

where R_q is the binary quadratic scoring rule (Equation 5.2), and x_j is the signal report of peer agent j.

The shadow posterior $p'(\cdot|x_i, y_i)$ in Candidate SSPP might have entries that fall outside $[0, 1]$ but this is not a problem because the properties we need with respect to the expected loss of the quadratic scoring rule are still well-defined (see Definition 11, Theorem 3.2, and Corollary 3.3.).

5.4.2 Incentive Analysis

The subjective equilibrium concept and strict ex post subjective incentive compatibility extend in the natural way from Section 5.3.2. As in BSPP, each agent is best-responding to the strategy of every other agent given knowledge of her own belief type, and with strict uncertainty on the belief types of other agents. The only difference to the concepts used there is that an agent's second report is now a signal and not a belief.

From a game-theoretic point of view, the key difference between BSPP and Candidate SSPP is that in Candidate SSPP there is an interdependency between an agent's first and second report. Agent i's first report y_i has an influence on the payment that she will receive for her second report x_i through its effect on $p'(\cdot|x_i, y_i)$. This requires a careful incentive analysis involving two steps. In Lemma 5.3, we first establish the optimal signal report given an already reported, fixed signal prior. There are two cases (the third is symmetric): if the reported signal prior is in between the two possible posteriors, then it is optimal for the agent to report her signal truthfully. If the reported signal prior is lower than both possible signal posteriors, then always reporting signal 2, independent of the observed signal is optimal. (In the symmetric case where the reported signal prior is higher than both possible posteriors, always reporting signal 1 is optimal.) Proposition 5.4 then compares the expected utility of these two cases, and derives the condition that needs to hold such that a signal report in between the two possible posteriors followed by the truthful report of the signal has higher expected utility than reporting a signal prior that is lower than both possible posteriors followed by signal report 2 (or the symmetric case).

Lemma 5.3 is a slight generalization of Theorem 3.5 (p. 31), in that the three different cases are spelled out.

Lemma 5.3. *In Candidate SSPP, given signal prior report y_i and realized signal $S_i = s_i$, agent i's optimal signal report x_i (assuming peer agent j is truthful) depends on her prior signal belief report y_i as follows:*

1. *If $y_i(2) < p_i(2|s_i)$, then she has a strict preference to report $x_i = 2$.*

2. *If $y_i(2) = p_i(2|s_i)$, then she is indifferent between $x_i \in \{1, 2\}$.*

3. *If $y_i(2) > p_i(2|s_i)$, then she has a strict preference to report $x_i = 1$.*

In particular, if $p_i(2|1) < y_i(2) < p_i(2|2)$, then the truthful report $x_i := s_i$ is optimal.

Proof. Fix belief report y_i and true signal s_i, and assume peer agent j's signal report is truthful. From Corollary 3.3 it follows that agent i should report x_i that leads to a shadow posterior $p'\big(2|x_i, y_i\big)$ with minimal distance to agent i's true posterior $p_i(2|s_i) = \Pr_i(S_j = 2 \mid S_i = s_i)$. Consider two cases:

1. If $y_i(2) < p_i(2|s_i)$ then $y_i(2) - p_i(2|s_i) < 0$, and $|y_i(2) + \delta - p_i(2|s_i)| < |y_i(2) - \delta - p_i(2|s_i)|$ and $x_i = 2$ is strictly optimal. The case of $y_i(2) > p_i(2|s_i)$ is symmetric.

2. If $y_i(2) = p_i(2|s_i)$ then the distance is the same for either report, and so indifference.

This completes the proof. □

Lemma 5.3 gives a hint as to what we are looking for—conditions on δ and an agent's subjective belief model, such that $p_i(2|1) < y_i(2) < p_i(2|2)$, and the agent's signal report is strictly truthful.

Proposition 5.4. *In Candidate SSPP, if mechanism parameter $0 < \delta \leq 2\big(p_i(2|2) - p_i(2|1)\big)$, agent i's strict best response to a truthful signal report by peer agent j is to make signal prior report $y_i(2) = p_i(2) \cdot (1 - \delta) + \frac{\delta}{2}$ (with $y_i(1) = 1 - y_i(2)$) and truthful signal report $x_i = s_i$.*

Proof. First of all, let us constrain agent i's strategy to reporting the true signal $x_i = s_i$. Given this, the expected score for reporting $y_i(2)$ in stage 1 is:

$$\begin{aligned} U_{\text{truesignal}}\big(y_i(2)\big) = \quad & p_i(2)\left(2y_i(2) - y_i(2)^2\right) \\ & + (1 - p_i(2))\left(1 - y_i(2)^2\right) \\ & + p_i(2)\left(2(y_i(2) + \delta) - (y_i(2) + \delta)^2\right) \\ & + (1 - p_i(2))\left(1 - (y_i(2) - \delta)^2\right). \end{aligned}$$

Taking the derivative with respect to $y_i(2)$, and setting to zero, we obtain

$$\begin{aligned} \frac{\partial U_{\text{truesignal}}\big(y_i(2)\big)}{\partial y_i(2)} &= 2\big(p_i(2) - y_i(2) - 2\delta p_i(2) + \delta + p_i(2) - y_i(2)\big) = 0 \\ \Leftrightarrow \quad y_i(2) &= p_i(2) + \frac{\delta\big(1 - 2p_i(2)\big)}{2} = p_i(2) \cdot (1 - \delta) + \frac{\delta}{2}, \end{aligned}$$

and a maximum by checking second-order conditions.

It is reassuring to confirm that for $\delta \leq 2\big(p_i(2|2) - p_i(2|1)\big)$, we have $p_i(2|1) < y_i(2) < p_i(2|2)$, and thus not only a feasible report, i.e. in the $[0,1]$ bound, but also consistency with Lemma 5.3: first, observe that for $p_i(2) = 0.5$, it holds that $y_i(2) = 0.5$ for any δ, so that $y_i(2)$ is in bounds. Second, observe that for $p_i(2) > 0.5$, it holds that $y_i(2) < p_i(2)$, and for $p_i(2) < 0.5$, it holds that $y_i(2) > p_i(2)$. Because of symmetry, it is then sufficient to show that $y_i(2) < p_i(2|2)$ for $p_i(2) < 0.5$, and thus $p_i(2|1) < p_i(2) < y_i(2) < p_i(2|2)$. We have,

$$\begin{aligned}
y_i(2) &= p_i(2) + \frac{\delta(1 - 2p_i(2))}{2} \leq p_i(2) + \frac{2\big(p_i(2|2) - p_i(2|1)\big)\big(1 - 2p_i(2)\big)}{2} \\
&= p_i(2) + \big(p_i(2|2) - p_i(2|1)\big)\big(1 - 2p_i(2)\big) \\
&= p_i(2) + p_i(2|2) - 2p_i(2)p_i(2|2) - p_i(2|1) + 2p_i(2)p_i(2|1) \\
&= p_i(2) - p_i(2)p_i(2|2) - p_i(2|1) + p_i(2)p_i(2|1) + p_i(2|2) \\
&\quad + p_i(2)p_i(2|1) - p_i(2)p_i(2|2) \\
&= p_i(2|2) + p_i(2)p_i(2|1) - p_i(2)p_i(2|2) < p_i(2|2).
\end{aligned}$$

For the final equality, we need that $p_i(2) = p_i(2)p_i(2|2) + \big(1 - p_i(2)\big)p_i(2|1)$ and thus $p_i(2) - p_i(2)p_i(2|2) - p_i(2|1) + p_i(2)p_i(2|1) = 0$, and the strict inequality follows from $p_i(2) > 0$ and $p_i(2|2) > p_i(2|1)$, which holds when stochastic relevance holds (Lemma 5.1).

What other strategy might be better for the agent? We know from Lemma 5.3 that the only other case to consider is $y_i(2) \leq p_i(2|1)$ (or symmetrically, $y_i(2) \geq p_i(2|2)$), where the optimal signal report is $x_i = 2$, independent of realized signal s_i. Given this, let us now constrain agent i's strategy to always reporting 2. Given this, the expected score for reporting $y_i(2)$ in stage 1 is

$$\begin{aligned}
U_{\text{alwayshigh}}\big(y_i(2)\big) = \quad & p_i(2)\big(2y_i(2) - y_i(2)^2\big) \\
& + \big(1 - p_i(2)\big)\big(1 - y_i(2)^2\big) \\
& + p_i(2)\Big(2\big(y_i(2) + \delta\big) - \big(y_i(2) + \delta\big)^2\Big) \\
& + \big(1 - p_i(2)\big)\big(1 - (y_i(2) + \delta)^2\big).
\end{aligned}$$

Taking the derivative with respect to $y_i(2)$, and setting to zero, we obtain

$$\frac{\partial U_{\text{alwayshigh}}\big(y_i(2)\big)}{\partial y_i(2)} = 2\big(p_i(2) - y_i(2) - \delta + p_i(2) - y_i(2)\big) = 0 \Leftrightarrow \;\; y_i(2) = p_i(2) - \frac{\delta}{2},$$

and a maximum by checking the second-order conditions. However, Candidate SSPP insists on $y_i \in \mathcal{D}$, and thus $y_i(2) \in [0,1]$. For $\delta > 2p_i(2)$ this is not feasible. Therefore, the expected utility given $y_i(2) = p_i(2) - \delta/2$ is an upper

bound on the actual utility available when playing the "always-high" strategy in stage 2.

Continuing, we establish that the expected loss, relative to being able to report $y_i(2) = p_i(2)$ in stage 1 and $p'(2|s_i) = p_i(2|s_i)$ in stage 2 is greater under the "always-high" strategy than the "true signal" strategy. The expected loss for the "true signal" strategy is:

$$L_{\text{truesignal}} = \left(\frac{\delta\big(1-2p_i(2)\big)}{2}\right)^2 + p_i(2)\left(p_i(2|2) - \left[p_i(2)\cdot(1-\delta) + \frac{\delta}{2} + \delta\right]\right)^2 + \big(1-p_i(2)\big)\left(p_i(2|1) - \left[p_i(2)\cdot(1-\delta) + \frac{\delta}{2} - \delta\right]\right)^2$$

For the always high strategy, the expected loss is:

$$L_{\text{alwayshigh}} \geq \left(\frac{\delta}{2}\right)^2 + p_i(2)\left(p_i(2|2) - \left[p_i(2) - \frac{\delta}{2} + \delta\right]\right)^2 + \big(1-p_i(2)\big)\left(p_i(2|1) - \left[p_i(2) - \frac{\delta}{2} + \delta\right]\right)^2$$

This is a lower bound on loss, because the optimal $y_i(2)$ in this case may be out of the $[0,1]$ bound and thus the agent's utility is upper-bounded by assuming $y_i(2) = p_i(2) - \delta/2$ is feasible. Combining, we have:

$$\begin{aligned}
L_{\text{alwayshigh}} - L_{\text{truesignal}} \geq & \left(\frac{\delta}{2}\right)^2 - \left(\frac{\delta\big(1-2p_i(2)\big)}{2}\right)^2 \\
& + p_i(2)\Bigg(\left(p_i(2|2) - \left[p_i(2) + \frac{\delta}{2}\right]\right)^2 \\
& \qquad - \left(p_i(2|2) - \left[p_i(2)\cdot(1-\delta) + \frac{3\delta}{2}\right]\right)^2\Bigg) \\
& + \big(1-p_i(2)\big)\Bigg(\left(p_i(2|1) - \left[p_i(2) + \frac{\delta}{2}\right]\right)^2 \\
& \qquad - \left(p_i(2|1) - \left[p_i(2)\cdot(1-\delta) - \frac{\delta}{2}\right]\right)^2\Bigg) \\
= & \delta^2 p_i(2)\cdot\big(1-p_i(2)\big) \\
& + \delta\big(1-p_i(2)\big)p_i(2)\Big(\big(-\delta + p_i(2|2) - p_i(2)\big) \\
& \qquad + \delta\big(p_i(2)-1\big) + p_i(2|2) - p_i(2)\Big) \\
& - \delta\big(1-p_i(2)\big)\big(1+p_i(2)\big)\big(\quad p_i(2|1) - p_i(2) + \delta p_i(2) \\
& \qquad + p_i(2|1) - p_i(2)\big)
\end{aligned}$$

$$= 2\delta\big(1 - p_i(2)\big)\big(\quad p_i(2) - p_i(2|1) - \delta p_i(2) \\ + p_i(2|2)p_i(2) - p_i(2|1)p_i(2)\big).$$

Since $2\delta\big(1 - p_i(2)\big) > 0$, we have

$$\begin{aligned} & L_{\text{alwayshigh}} - L_{\text{truesignal}} > 0 \\ \Leftrightarrow\ & p_i(2) - p_i(2|1) - \delta p_i(2) + p_i(2|2)p_i(2) - p_i(2|1)p_i(2) \geq 0 \\ \Leftrightarrow\ & \big(p_i(2) - p_i(2|1)\big) + p_i(2)\big(p_i(2|2) - p_i(2|1)\big) \geq p_i(2)\delta \\ \Leftrightarrow\ & p_i(2)\big(p_i(2|2) - p_i(2|1)\big) + p_i(2)\big(p_i(2|2) - p_i(2|1)\big) \geq p_i(2)\delta \\ \Leftrightarrow\ & \delta \leq 2\big(p_i(2|2) - p_i(2|1)\big). \end{aligned}$$

This establishes that a prior signal belief report $y_i(2) = p_i(2) + \frac{\delta\left(1-2p_i(2)\right)}{2}$ and the truthful signal report constitute a strict best response to a truthful signal report by agent j. □

We immediately have the following:

Theorem 5.5. *In a strict ex post subjective equilibrium of Candidate SSPP, and given temporal separation, every agent i reports her signal truthfully if mechanism parameter $0 < \delta \leq 2\big(p_i(2|2) - p_i(2|1)\big)$ for all $i \in \{1, \ldots, n\}$ and given stochastic relevance.*

This provides partial incentive compatibility: an agent has strict incentives to report her true signal in stage 2 but should deviate to $y_i(2) = p_i(2) + \frac{\delta(1-2p_i(2))}{2}$ (with $y_i(1) = 1 - y_i(2)$) in stage 1. The only informational requirement on the mechanism is that it must pick a δ small enough, depending on knowledge of a valid $\epsilon > 0$ such that $\epsilon \leq p_i(2|2) - p_i(2|1)$ for all $i \in \{1, \ldots, n\}$ (with $\delta \leq 2\epsilon$ being sufficient). Certainly, stochastic relevance implies $p_i(2|1) < p_i(2) < p_i(2|2)$ for all belief types. What is required in addition is knowledge of this minimal "degree of informativeness" of a signal, valid for every belief type in the population.

5.5 Shadow Subjective-Prior Mechanism (SSPP)

We can now apply the revelation principle and achieve strict incentive compatibility in regard to both the belief report and the signal report. The crucial observation is that the optimal misreport $y_i(2) = p_i(2) + \frac{\delta\left(1-2p_i(2)\right)}{2}$ (with $y_i(1) = 1 - y_i(2)$) depends *only* on the agent's signal prior p_i and parameter δ of the mechanism. It does not depend on any other aspect of the agent's belief type. For this reason, the mechanism can simply compute y_i directly on behalf of the agents on the basis of p_i.

5.5.1 Mechanism

The *shadow subjective-prior mechanism (SSPP)* is defined as:

1. (Stage 1) Ask agent i for her signal prior report $y_i \in \mathcal{D}$.
2. Agent i observes signal $S_i = s_i$.
3. (Stage 2) Ask agent i for her signal report $x_i \in \{1, 2\}$, and calculate shadow posterior

$$p'(\cdot|x_i, y_i) = \begin{cases} \begin{pmatrix} y_i(1) - \frac{\delta\left(1-2y_i(2)\right)}{2} + \delta \\ y_i(2) + \frac{\delta\left(1-2y_i(2)\right)}{2} - \delta \end{pmatrix} & \text{if} \quad x_i = 1 \\ \begin{pmatrix} y_i(1) - \frac{\delta\left(1-2y_i(2)\right)}{2} - \delta \\ y_i(2) + \frac{\delta\left(1-2y_i(2)\right)}{2} + \delta \end{pmatrix} & \text{if} \quad x_i = 2, \end{cases}$$

   where $\delta > 0$ is a parameter of the mechanism.

4. For each agent i, choose peer agent $j = i + 1 \,(\text{modulo } n)$ and pay agent i:

$$u_i\big(y_i, x_i, x_j\big) = R_q\left(y_i(2) + \frac{\delta\big(1 - 2y_i(2)\big)}{2}, x_j\right) + R_q\Big(p'\big(\cdot|x_i, y_i\big), x_j\Big)$$

   where R_q is the binary quadratic scoring rule (Equation 5.2), and x_j is the signal report of peer agent j.

Observe that, for $\delta = 1$, the transformed signal prior report $y_i(2) + \frac{\delta\left(1-2y_i(2)\right)}{2}$ becomes $\frac{1}{2}$ which does not depend on $y_i(2)$. This would mean that while agent i still cannot do better than reporting $y_i = p_i$, the signal prior report would not be *strictly* incentivized. Excluding this case, the following theorem then follows immediately from Proposition 5.4 and the equilibrium analysis of Candidate SSPP:

Theorem 5.6. *SSPP is strictly ex post subjective incentive compatible given temporal separation, if mechanism parameter* $\delta \neq 1$ *and* $0 < \delta \leq 2\big(p_i(2|2) - p_i(2|1)\big)$ *for all* $i \in \{1, \ldots, n\}$.

5.5.2 Compact SSPP

In SSPP, the signal prior report y_i affects both parts of agent i's score while the signal report x_i only affects the second part. This raises the question as to whether the second part alone could suffice for strict truth-telling incentives. That is, whether Candidate SSPP's payment rule can be shortened to $R_q\Big(p'\big(\cdot|x_i, y_i\big), x_j\Big)$.

In fact, this is possible. First observe that Lemma 5.3 still holds since this analysis pertained only to the score for the signal report. Given this, the outline of the analysis follows as before. We can derive (a) the optimal report $y_i(2)$ given that the agent reports her true signal s_i, and (b) the optimal report $y_i(2)$ given that the agent always reports signal 2 (or symmetrically, always reports signal 1.) Ignoring the symmetric case of "always low", this yields:

1. If $x_i = 2$ (independent of s_i), the optimal belief report is $y_i(2) = p_i(2) - \delta$.

2. If $x_i = s_i$, the optimal belief report is $y_i(2) = p_i(2) \cdot (1 - 2\delta) + \delta$.

Considering the expected loss relative to being able to make perfect reports $p_i(2|1)$ or $p_i(2|2)$ in stage 2 (depending on the observed signal), and requiring that the loss from reporting the true signal and $y_i(2) = p_i(2) \cdot (1 - 2\delta) + \delta$ is strictly less than that of always reporting signal 2 and $y_i(2) = p_i(2) - \delta$, the constraint on parameter δ is $0 < \delta \leq p_i(2|2) - p_i(2|1)$.[2] Given this, and adopting the revelation principle as before, we obtain the compact version of SSPP.

The *compact shadow subjective-prior mechanism (Compact SSPP)* is defined as:

1. (Stage 1) Ask agent i for her signal prior report $y_i \in \mathcal{D}$.

2. Agent i observes signal $S_i = s_i$.

3. (Stage 2) Ask agent i for her signal report $x_i \in \{1, 2\}$, choose peer agent $j = i + 1$ (modulo n) and pay agent i:

$$u_i\big(x_i, y_i, x_j\big) = R_q\big((1 - 2\delta) \cdot y_i(2) + 2\delta(x_i - 1), x_j\big), \tag{5.3}$$

   where $\delta > 0$ is a parameter of the mechanism, R_q is the binary quadratic scoring rule (Equation 5.2), and x_j is the signal report of peer agent j.

Analogous to SSPP, the transformed signal prior report in Compact SSPP is independent of $y_i(2)$ if $\delta = \frac{1}{2}$, so that, in that case, the signal prior report $y_i(2)$ would only be weakly truthful. I thus exclude $\delta = \frac{1}{2}$ in the following theorem, which I state without proof:

Theorem 5.7. *Compact SSPP is strictly ex post subjective incentive compatible given temporal separation if parameter $\delta \neq \frac{1}{2}$ and $0 < \delta \leq p_i(2|2) - p_i(2|1)$ for all $i \in \{1, \ldots, n\}$.*

[2] For $0 < \delta \leq p_i(2|2) - p_i(2|1)$, one can again confirm that $y_i(2) = p_i(2) \cdot (1 - 2\delta) + \delta$ is a feasible report. As in the analysis of SSPP, it is irrelevant that $y_i(2) = p_i(2) - \delta$ may not be in the $[0, 1]$ range since requiring a report inside the range can only further reduce an agent's utility.

The payment rule of Compact SSPP has a nice intuitive interpretation. It can be written as $R_q\big((1-\eta)\cdot y_i(2)+\eta(x_i-1), x_j\big)$ with $\eta = 2\delta$. Observe that the expected value of $x_i - 1$ is $p_i(2)$. Therefore, given that agent i is truthful, the expected belief report that is applied to the quadratic scoring rule is $\mathbf{E}\big[(1-\eta)\cdot y_i(2)+\eta(x_i-1)\big] = p_i(2)$. This makes sense given that the quadratic scoring rule is strictly proper, and $p_i(2)$ is agent i's best prediction for the signal report x_j at stage 1 given that agent j is truthful. The role of $\eta > 0$ is to put some weight on the signal report, so that the agent has an incentive to use the signal report in obtaining better accuracy in regard to her signal posteriors.

On the other hand, if η gets too large relative to the effect of an agent's signal on her signal posterior, then the agent prefers not to adjust her shadow posterior through reporting a signal in a way that depends on the observed signal. Instead, in this situation where the two possible signal posteriors are relatively close to each other, she will set $\mathbf{E}\big[(1-\eta)\cdot y_i(2)+\eta(x_i-1)\big] = p_i(2)$ by choosing $x_i = 2$ (always) and $y_i(2) = \frac{p_i(2)-\eta}{1-\eta} = \frac{p_i(2)-2\delta}{1-2\delta}$. By applying this to the transformation due to the direct-revelation approach, one rediscovers the optimal misreport of Candidate SSPP given a high signal report $\frac{p_i(2)-2\delta}{1-2\delta}(1-2\delta)+\delta = p_i(2)-\delta$.

An interesting special case is choosing $\delta = 0.5$, which is a valid parametarization for *weakly truthful* signal prior incentives when $0.5 \leq p_i(2|2) - p_i(2|1)$ and thus $p_i(2|2) > 0.5$ and $p_i(2|1) < 0.5$. In this case, the effect of belief report $y_i(2)$ disappears and the shadow belief report that is adopted in the scoring rule is 1 if the signal report is 2 and 0 if the signal report is 1. This coincides with simple output agreement (Section 2.4), where agent i obtains a payment of 1 if her signal report agrees with that of agent j and 0 otherwise. It makes sense that this would retain strict incentives in regard to the signal report in this case, since $p_i(2|2) > 0.5$ and $p_i(2|1) < 0.5$, and so shadow posteriors 1 and 0 are minimizing the distances to $p_i(2|2)$ and $p_i(2|1)$, respectively.

5.5.3 Individual Rationality

In the BSPP mechanism from Section 5.3, which utilizes two belief reports, using the normalized quadratic scoring rule R_q giving scores between 0 and 1 ensures ex post individual rationality. For the SSPP mechanisms from Sections 5.4 to 5.5.2, it is no longer obvious that this still holds because there could be out-of-bound shadow posterior reports (outside of $[0,1]$), so that the *ex post* scores may be negative. For Compact SSPP and $\delta > 0.5$, this can indeed be the case[3] but since $\delta > 0.5$ implies $p_i(2|2) > 0.5$ and $p_i(2|1) < 0.5$, simple output

[3]The lowest possible score is attained through the lowest possible $(1-2\delta)\cdot y_i(2)+2\delta x_i$ and $x_j = 2$, or through the highest possible $(1-2\delta)\cdot y_i(2)+2\delta x_i$ and $x_j = 1$. For $\delta > 0.5$, $(1-2\delta)\cdot y_i(2)+2\delta x_i$ is minimized by reporting $y_i = 1$ and $x_i = 1$, and maximized by reporting $y_i = 0$ and $x_i = 2$. Applied to R_q together with $x_j = 1$ and $x_j = 1$, respectively, one obtains

agreement is already strictly truthful without requiring an additional report, so that only $\delta < 0.5$ is meaningful in Compact SSPP.

For $\delta \leq 0.5$, Compact SSPP's score is already in between 0 and 1. To see this, consider the range of possible values for $(1-2\delta)\cdot y_i(2) + 2\delta x_i$: the signal report x_i is either 1 or 2, and so this becomes either $(1-2\delta)\cdot y_i(2)$ or $(1-2\delta)\cdot y_i(2) + 2\delta$, with the latter being strictly larger than the former. For $\delta \leq 0.5$, then $(1-2\delta)\cdot y_i(2) \geq 0$ and $(1-2\delta)\cdot y_i(2) + 2\delta \leq 1$ for all $y_i(2) \in [0,1]$. Moreover, 0 and 1 is obtained by reporting $\big(y_i(2) = 0, x_i = 1\big)$ and $\big(y_i(2) = 1, x_i = 2\big)$, respectively, so that for $\delta \leq 0.5$, Compact SSPP attains scores between 0 and 1.

Example 9. *As in Example 8 for BSPP, consider again the ad exchange example from Section 2.2 with numbers* $m = 2$, $\Pr_i(T = 2) = 0.3$, $\Pr_i(S = 2 \mid T = 2) = 0.6$, *and* $\Pr_i(S = 2 \mid T = 1) = 0.1$. *The procedure using Compact SSPP with* $\delta = 0.1$, *such that* $\delta \leq p_i(2|2) - p_i(2|1) = 0.28$ *and* $\delta \leq 0.5$ *(for individual rationality) is:*

1. *Worker i accepts the task that was posted on Amazon Mechanical Turk by the ad exchange. She is asked for her signal prior report of observing violence, and she truthfully reports*

$$y_i(2) = 0.25.$$

2. *A website that may or may not contain violent content is shown to worker i and, after looking at it carefully, she is asked to report if she observed violence on the website. She did not observe violence, i.e. $S_i = 1$, and so she updates her signal posterior to $p_i(2|1) = 0.18$, and truthfully reports*

$$x_i = 1.$$

3. *Another worker $j \neq i$ also follows Steps 1 and 2, with potentially different beliefs and experiences.*

4. *Worker i is scored against worker j's reported signal x_j. Assuming agent j also did not observe violent content on the website and reported this truthfully, i.e. $x_j = 1$, agent i is paid*

$$R_q\big(0.8 \cdot 0.25 + 0.2 \cdot 0, 1\big) = R_q\big(0.2, 1\big) = 0.96$$

the lowest possible score as $2(1-2\delta) - (1-2\delta)^2 = 1 - (2\delta)^2 = 1 - 4\delta^2$. Therefore, by adding $|1 - 4\delta^2| = 4\delta^2 - 1$ to every agent's score, Compact SSPP can be made individually rational in the $\delta > 0.5$ case.

5.6 Conclusion

In this chapter, I presented two incentive compatible mechanisms for the elicitation of truthful user feedback that escape the strong common knowledge assumptions of the peer prediction mechanisms from Chapters 2 to 4. I believe that this development is of significant practical importance. The *compact shadow subjective-prior mechanism (Compact SSPP)* from Section 5.5.2 provides a particularly simple intuitive interpretation, is easy to analyze, and aligns incentives with truthful reporting of the signal prior and the signal. To the best of my knowledge, in terms of belief structure, the setting I study is the most general that has been studied in the context of peer prediction mechanisms. The theoretical analysis adopts a solution concept that is weaker than dominant strategy equilibrium but stronger than Bayes-Nash equilibrium.

Discussion: Application

I offer some remarks in regard to considerations when making use of SSPP mechanisms in practical applications.

Information aggregation

In the classical peer prediction method, the mechanism uses the common belief model to compute the signal posteriors and publish this information. In our setting with private and subjective belief models, the only objective information, i.e. the only information stemming from the world, are the signals. The mechanism can therefore simply publish the percentage of positive signal reports, allowing each agent to incorporate this information into her own subjective belief model. In fact, this is common practice. Hotwire,[4] for example, publishes the percentage of customers who have reported that they were satisfied with their stay at a given hotel. Another example is eBay,[5] which publishes the percentage of positive reports for a given seller.

User interface (UI)

The SSPP mechanisms require users to make reports about probabilistic beliefs. While it would be difficult to design a UI that makes reporting full distributions user friendly, I believe there are UIs that can achieve this for the binary setting, in which only a single probability is required. A suitable user interface "hides" the probabilities from users by adopting a point scale from 0 to 10. These points would directly correspond to probabilities, and allow users to interact

[4] `http://www.hotwire.com`
[5] `http://www.ebay.com`

with the system in a way they are familiar with from other online rating sites (albeit introducing a forced approximation to their reports). Of course, the right choice of UI depends on the application, and probability reports may sometimes be feasible. For example, when booking a hotel on Hotwire, the following question seems reasonably easy to answer: *"What is your prediction that another customer will recommend this hotel to a friend?"*

The shadow subjective-prior mechanisms (SSPP and Compact SSPP) have the same type of reports as Bayesian Truth Serum mechanisms (Chapter 4): a prediction report about the experiences of other agents and a signal report. In a study with inexpert human raters on Amazon Mechanical Turk, it has been shown that human agents can indeed report such information successfully [Shaw et al., 2011]. In fact, in this experimental comparison of different incentive schemes, the (original) Bayesian Truth Serum (Section 4.3) elicited the responses with highest quality among all tested schemes.

Future Work

An interesting direction for future work is to generalize the mechanisms of this chapter to multiple signals. For the BSPP mechanism, this will be a simple application of the multi-signal Shadowing Method (Section 3.6). For the more complex shadowing subjective-prior mechanisms, instead of re-doing the analysis for multiple signals from candidate mechanism (Section 5.4) to compact mechanism (Section 5.5.2), it will be interesting whether one can generalize the simple form of Compact SSPP directly.

Another interesting direction for future work is see whether other domains of mechanism design allow for ex post subjective incentive compatible mechanisms. A natural first candidate would be mechanisms with correlated types, e.g. a private and subjective version of the full surplus extraction result in the style of Crémer and McLean [1985, 1988].

Chapter 6

Minimal-Reporting Subjective-Prior Peer Prediction

The classical peer prediction method (Chapter 2) is a strictly truthful peer prediction mechanism but it assumes all agents share the same belief model and that the mechanism knows this model. Bayesian Truth Serum mechanisms (Chapter 4) relax the latter requirement that the mechanism knows the common belief model at the cost of "non-minimality," i.e., users need to report both their signals and a belief about the signals of others. Subjective Peer Prediction mechanisms (Chapter 5) further relax the assumption on the common knowledge, in that agents can have subjective beliefs with regard to the belief model, which the mechanism does not need to know. However, neither the basic subjective-prior mechanism (BSPP, Section 5.3) nor the shadow subjective-prior mechanisms (SSPP and Compact SSPP, Section 5.5) are minimal. BSPP requires two belief reports, and SSPP requires a belief and a signal report, the same type of reports as Bayesian Truth Serum mechanisms. Other methods, such as the Shadowing Method (Chapter 3), insist on minimality but still require knowledge of the signal prior. This suggests a trade-off between the robustness of incentive properties and the reporting requirements. In this chapter, I further explore this trade-off and develop the theoretical foundation for *learning* the signal prior in combination with minimal peer-prediction methods.

The main difference between this chapter's model and the model from Chapter 5 (which already incorporates subjective belief models) is that agents report on several different items or tasks. For example, guests may report whether they would recommend each of multiple hotels to a friend. Similarly, following the

example of Section 2.2, there could be multiple websites, with crowd workers asked to report whether they contain offensive content.

I will refer to a particular hotel or website as an *item*. It is important to note that just as in the peer prediction mechanisms of Chapters 2 to 5, any single agent only needs to report on one item on which at least one other agent submits a report. What is required on top of this is that there are other agents reporting on other items as well. The reason for this additional requirement is that, to score an agent, the mechanism will use the empirical distribution of reported signals on items the agent has not seen or reported on as the "empirical signal prior," to which the mechanism then applies the Shadowing Method (Section 3.5).

The main result is a strictly truthful mechanism that allows agents to have subjective belief models. All that is assumed to be common knowledge is a lower bound on the extent of the belief change from signal prior to signal posterior. The mechanism delays payments until the empirical distribution of signals is accurate enough, and I provide a bound on the number of items required for the mechanism to provide strict incentives for truthful reporting. Moreover, I insist on *minimality*, i.e. agents are only asked to report their signals. This mechanism is the first strictly truthful mechanism that combines minimality with subjective belief models.

The remainder of this chapter is organized as follows. In Section 6.1, I discuss related work particular to this chapter. In Section 6.2, I explain the difference of this chapter's model as compared to the subjective model used in the preceding chapter (Section 5.2). In Section 6.3, I then introduce the *Empirical Shadowing Method*, which allows agents to have subjective belief models as, and where the signal prior required for the Shadowing Method (Chapter 3) is learned from other agents' signal reports. Section 6.4 shows how to use a form of Hoeffding's inequality to derive upper bounds on the required number of samples given a lower bound on the extent of the belief change from signal prior to signal posterior. I conclude with a discussion and interesting directions for future work in Section 6.5.

6.1 Related Work

In addition to the related work on peer prediction mechanisms introduced in Chapters 2 to 5, there is additional related work particularly relevant to the mechanism presented in this chapter.

Jurca and Faltings [2008, 2011] suggest a mechanism for *on-line* opinion polls, which is situated in the same setting with commonly-held belief model as Bayesian Truth Serum mechanisms (Chapter 4). The mechanism is minimal,

i.e. it requires only a signal report, and it publishes the empirical frequencies of reports, which, in equilibrium, converge towards the true distribution of signals in the population. A building block of their mechanism is the 1/prior mechanism (Section 3.3).

The work in this chapter is different from the work by Jurca and Faltings in several aspects. First, the on-line polling setting is orthogonal to the setting of this chapter in that with opinion polls, there is one item with many reports about that one item, whereas in this chapter, we have many items with few reports for each item. Second, the opinion poll mechanism is not truthful, i.e. agents cannot simply report their true signal. Instead, agents need to choose their signal report depending on the current empirical frequency of past reports, resulting in higher cognitive costs on behalf of the agents. The mechanism presented in this chapter is truthful, resulting in low cognitive cost as agents can simply report their true signal. Third, the opinion poll mechanism is restricted to the common-prior model of Chapter 4, whereas the mechanism in this chapter allows for subjective belief models.

Jurca and Faltings also prove an impossibility result in regard to achieving strict truthfulness with minimal reporting and a belief model that is unknown to the mechanism. It does not apply to the mechanism presented in this chapter because I assume common knowledge of a lower bound on the extent of the belief change from prior to posterior.

6.2 Model

The model of this chapter is a variation of the binary-signal model with private and subjective belief models (Section 5.2), where there is not only one item (e.g. one hotel, product, or task) that all agents experience, but $g > 1$ items, each of which is experienced by at least two agents. To keep the presentation simple, I associate each agent with a single item and assume that there are two agents per item. Each agent i is indexed such that agent i belongs to item $i \,(\text{modulo}\, g)$. This is for presentational reasons only. For example, the modification to allow each agent to experience multiple items just requires care to only use items the agent did not experience when calculating the empirical frequency with which to score that agent. Note that we cannot use any reports from items the agent experienced (not even reports by other agents) because, through her experience with that item, the agent learns something about the instantiated state of the item, and, from the agent's perspective, signals would thus not be drawn according to the signal prior distribution anymore.

With regard to agent i's subjective belief model, the assumption is that signals of items that agent i has not experienced are sampled according to her

subjective signal prior $p_i(\cdot)$. Her subjective belief that an agent observes a high signal from an item she has not experienced is thus $p_i(2)$. It is important to emphasize that given this model, agent i observing a signal from her own item changes her belief about the signals observed by other agents experiencing the same item, whereas $p_i(2)$ remains agent i's belief about the distribution of high signals over all other items.

Furthermore, it is important to emphasize that an agent's subjective belief model reflects her beliefs about the behavior of an item, which affects all agents in the same way. That is, as in all preceding chapters, agent i is blind to the identity of an agent and does not distinguish between her belief that agent j receives a high signal and her belief that some other agent $k \neq j$ receives a high signal. As in Chapter 5, one can thus adopt shorthand $p_i(2|s_i) = \Pr_i(S_j = 2 \mid S_i = s_i)$ for agent i's signal posterior that any other agent j experiencing the same item receives a high signal given agent i's signal s_i.

Moreover, it is assumed that the signal posteriors are fully mixed, i.e. $p_i(s|s') > 0$ for all $s, s' \in \{1, 2\}$. In terms of basic model parameters, this condition is, for example, satisfied if $\Pr_i(S = s|T = t) > 0$ for all $s \in \{1, 2\}, t \in \{1, \ldots, l\}$, and $i \in \{1, \ldots, n\}$. Lemma 6.1 follows immediately from combining Lemma 5.1 with the assumption of strictly mixed signal posteriors.

Lemma 6.1. *Given stochastic relevance of θ_i, it holds that $1 > p_i(2|2) > p_i(2) > p_i(2|1) > 0$.*

The main requirement on the knowledge of the mechanism and the agents is that there is a common-knowledge lower bound $\lambda > 0$ on the distance between the signal prior $p_i(2)$ and the signal posteriors $p_i(2|1)$ and $p_i(2|2)$, respectively, i.e.

$$\lambda \leq \min\big(p_i(2) - p_i(2|1), p_i(2|2) - p_i(2)\big). \tag{6.1}$$

This provides a lower bound on how much the belief of an agent changes through observing a signal. It can be chosen to be arbitrarily small as long as it is strictly positive. The main result is a statement about the trade-off between a low belief change bound λ and a low number of required items, i.e. for any bound λ, we will obtain a number g, such that if there are at least g items, the mechanism is strictly truthful.

6.3 The Empirical Shadowing Method

We are now ready to define an approach that adopts the empirical distribution of signal reports in place of the mechanism's assumed knowledge of the

signal prior in the Shadowing Method (Section 3). This Empirical Shadowing Method is minimal, i.e. it elicits reports consisting of only signals. Note that the mechanism withholds payments until every agent has reported her signal.

The *Empirical Shadowing Method* is defined as:

1. Ask agent i for her signal report $x_i \in \{1, 2\}$.

2. Let N_i be a set of agents, such that there is one agent associated with every of the g items except the item associated with agent i. Compute the empirical frequency (empirical mean) $\hat{p}_i(2)$ of all $g-1$ signal reports of agents in N_i:

$$\hat{p}_i(2) = \sum_{k \in N_i} \frac{x_k}{g-1}.$$

3. Using this empirical frequency of high signals $\hat{p}_i(2)$ and agent i's signal report x_i, calculate shadow posterior

$$p'\big(2|x_i, \hat{p}_i(2)\big) = \begin{cases} \hat{p}_i(2) + \delta, & \text{if} \quad x_i = 2 \\ \hat{p}_i(2) - \delta, & \text{if} \quad x_i = 1 \end{cases}, \tag{6.2}$$

   where $\delta > 0$ is a parameter of the mechanism.

4. For each agent i, choose a peer agent j that is experiencing the same item as agent i, and pay agent i:

$$u_i(x_i, \hat{p}_i(2), x_j) = R_q\Big(p'\big(2|x_i, \hat{p}_i(2)\big), x_j\Big)$$

   where R_q is the quadratic scoring rule and x_j is the signal report of peer agent j.

6.4 Incentive Analysis

There are several ways of analyzing the incentives of the Empirical Shadowing Method. I present a first approach based on establishing a lower bound on the probability that the empirical frequency of high signals, $\hat{p}_i(2)$, lies "in between" agent i's two possible signal posteriors $p_i(2|1)$ and $p_i(2|2)$, so that the Shadowing Method is strictly truthful (Corollary 3.6).

Of course, with a finite number of items, there is always some probability that $\hat{p}_i(2) \leq p_i(2|1)$ or $\hat{p}_i(2) \geq p_i(2|2)$, in which case the Shadowing Method is not truthful. I derive a lower bound on the expected benefit of being truthful given that $\hat{p}_i(2)$ lies within the interval and an upper bound on the expected

benefit from a misreport when $\hat{p}_i(2)$ is outside the interval. Together with an upper bound on the probability that the empirical frequency is outside the interval, this provides a bound on the number of items required for the Empirical Shadowing Method mechanism to have strict incentives for agents to be truthful.

I begin with two technical lemmas.

Lemma 6.2. *Given stochastic relevance of θ_i, it holds that $p_i(2) - p_i(2|1) \leq p_i(2|2) - p_i(2) \Leftrightarrow p_i(2) \leq 0.5$.*

Proof. Note that

$$\begin{aligned} p_i(2|2) =& 1 - p_i(1|2) = 1 - \frac{p_i(1)}{p_i(2)} p_i(2|1) = 1 - \frac{1 - p_i(2)}{p_i(2)} p_i(2|1) \\ =& 1 - \frac{p_i(2|1)}{p_i(2)} + p_i(2|1). \end{aligned}$$

So we have

$$\begin{aligned} & p_i(2) - p_i(2|1) \leq p_i(2|2) - p_i(2) \\ \Leftrightarrow & p_i(2) - p_i(2|1) \leq 1 - \frac{p_i(2|1)}{p_i(2)} + p_i(2|1) - p_i(2) \\ \Leftrightarrow & 2p_i(2) - 2p_i(2|1) + \frac{p_i(2|1)}{p_i(2)} - 1 \leq 0 \\ \Leftrightarrow & 2p_i(2)^2 - 2p_i(2|1)p_i(2) + p_i(2|1) - p_i(2) \leq 0 \\ \Leftrightarrow & \big(2p_i(2) - 1\big) \underbrace{\big(p_i(2) - p_i(2|1)\big)}_{>0 \text{ (Lemma 6.1)}} \leq 0 \\ \Leftrightarrow & p_i(2) \leq 0.5. \end{aligned}$$

This completes the proof. □

Lemma 6.3. *The smallest possible $p_i(2|2)$ given belief change bound λ and stochastic relevance is:*

$$\underline{p}_i(2|2) = \begin{cases} 2\sqrt{\lambda} - \lambda, & \text{if} \quad p_i(2) \leq 0.5 \\ 0.5 + \lambda, & \text{if} \quad p_i(2) \geq 0.5, \end{cases} \tag{6.3}$$

A lower bound for $p_i(2|2)$ given belief change bound λ is $2\sqrt{\lambda} - \lambda$.

Proof. I first prove the statement for $p_i(2) \geq 0.5$. From Lemma 6.2, we know that $\lambda \leq p_i(2|2) - p_i(2)$ entails $\lambda \leq p_i(2) - p_i(2|1)$, so that it is sufficient to minimize $p_i(2|2)$ subject to $\lambda \leq p_i(2|2) - p_i(2)$. Then $p_i(2|2) = p_i(2) + \lambda$ which is minimized for $p_i(2) = 0.5$, so that $\underline{p}_i(2|2) = 0.5 + \lambda$ if $p_i(2) \geq 0.5$.

For the case where $p_i(2) \leq 0.5$, we can restrict the analysis to $\lambda \leq p_i(2) - p_i(2|1)$ because we know from Lemma 6.2 that for $p_i(2) \leq 0.5$, $\lambda \leq p_i(2) - p_i(2|1)$

entails $\lambda \leq p_i(2|2) - p_i(2)$. From the proof of Lemma 6.2 we also know that

$$p_i(2|2) = 1 - \frac{p_i(2|1)}{p_i(2)} + p_i(2|1) = 1 - p_i(2|1)\left(\frac{1}{p_i(2)} - 1\right) \tag{6.4}$$

Equivalent to minimizing $p_i(2|2)$ given $\lambda \leq p_i(2) - p_i(2|1)$ is thus maximizing $p_i(2|1)\left(\frac{1}{p_i(2)} - 1\right)$ given $\lambda \leq p_i(2) - p_i(2|1)$. This is maximized for the largest possible $p_i(2|1)$ and the smallest possible $p_i(2)$, so that a necessary condition for a minimal $p_i(2|2)$ is $p_i(2|1) = p_i(2) - \lambda$. Using this in (6.4) we obtain:

$$\underline{p}_i(2|2) = 1 - \left(\frac{p_i(2) - \lambda}{p_i(2)} - \big(p_i(2) - \lambda\big)\right) = \frac{\lambda}{p_i(2)} + p_i(2) - \lambda \tag{6.5}$$

Taking the derivative and setting to 0 one obtains:

$$\frac{\partial \underline{p}_i(2|2)(p_i(2))}{\partial p_i(2)} = 1 - \frac{\lambda}{p_i(2)^2} = 0 \Leftrightarrow p_i(2) = \sqrt{\lambda}$$

Inserting this back into (6.5), one obtains the minimal $p_i(2|2)$ for $p_i(2) \leq 0.5$:

$$\underline{p}_i(2|2) = \frac{\lambda}{\sqrt{\lambda}} + \sqrt{\lambda} - \lambda = 2\sqrt{\lambda} - \lambda. \tag{6.6}$$

Since $2\sqrt{\lambda} - \lambda \leq 0.5 + \lambda$ for all $0 < \lambda < 0.5$, this completes the proof. □

The proof of Theorem 6.5 uses a form of Hoeffding's inequality in order to be able to make a statement about the number of items (samples) that are required without knowledge of $p_i(2)$. (For the simple steps showing how to get from the standard formulation to the formulation of Lemma 6.4, see, for example, p. 3 in Domke [2010].)

Lemma 6.4. *[Hoeffding, 1963] Let $Z_1, \ldots, Z_g \in [0,1]$ be independent and identically distributed random variables. If*

$$g \geq \frac{1}{2\epsilon^2} \ln\left(\frac{2}{d}\right),$$

for $\epsilon > 0$, $0 < d < 1$, then $\Pr(\left|\frac{1}{g}\sum_{i=1}^{g} Z_i - E[Z]\right| \leq \epsilon) \geq 1 - d$. That is, with probability at least $1 - d$, the difference between the empirical mean $\frac{1}{g}\sum_{i=1}^{g} Z_i$ and the expected value $E[Z]$ is at most ϵ.

Theorem 6.5. *The Empirical Shadowing Method is strictly ex post subjective incentive compatible given belief change bound λ and $g - 1$ samples with $g \geq \frac{1}{2\epsilon^2} \ln\left(\frac{2\left(1+2(\lambda-\sqrt{\lambda})-\epsilon\right)}{\lambda-\epsilon}\right) + 2$ and $0 < \epsilon < \lambda$.*

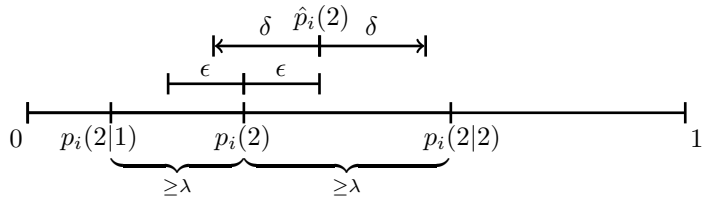


Figure 6.1: Illustration of Case 1 in the analysis of the Empirical Shadowing Method. Observe that $\epsilon < \lambda$, so that $\hat{p}_i(2) = p_i(2) + \epsilon < p_i(2|2)$. (Depending on parameter $\delta > 0$, it may or may not hold that $\hat{p}_i(2) + \delta < p_i(2|2)$.)

Proof. Given that all other agents are truthful, I show that agent i's unique best response is to be truthful. To apply Hoeffding's inequality, I introduce some $\epsilon > 0$, $\epsilon < \lambda$ and analyze two cases: the case where the empirical frequency is no more than ϵ away from the signal prior $p_i(2)$, i.e. $|\hat{p}_i(2) - p_i(2)| \leq \epsilon$, and the case where the empirical frequency is further away than ϵ, i.e. $|\hat{p}_i(2) - p_i(2)| > \epsilon$.

(Case 1): $|\hat{p}_i(2) - p_i(2)| \leq \epsilon$. (Also see Figure 6.1.)

From $0 < \epsilon < \lambda \leq \min(p_i(2) - p_i(2|1), p_i(2|2) - p_i(2))$ it follows that $p_i(2|1) < \hat{p}_i(2) < p_i(2|2)$, so that the shadowing method elicits signals truthfully. I proceed to quantify this positive expected benefit of reporting truthfully using the difference in expected loss given $S_i = 2$ (case $S_i = 1$ is analogous). Recall that the quadratic scoring rule has quadratic loss (Theorem 3.2):

$$
\begin{aligned}
&\Delta U_i(x_i = 2|S_i = 2) \\
=&U_i(x_i = 2|S_i = 2) - U_i(x_i = 1|S_i = 2) \\
=&-\left(p_i(2|2) - (\hat{p}_i(2) + \delta)\right)^2 + \left(p_i(2|2) - (\hat{p}_i(2) - \delta)\right)^2 \\
=&\Big(\big(p_i(2|2) - (\hat{p}_i(2) - \delta)\big) + \big(p_i(2|2) - (\hat{p}_i(2) + \delta)\big)\Big) \\
&\qquad \cdot \Big(\big(p_i(2|2) - (\hat{p}_i(2) - \delta)\big) - \big(p_i(2|2) - (\hat{p}_i(2) + \delta)\big)\Big) \\
=&\left(2p_i(2|2) - 2\hat{p}_i(2)\right) 2\delta = 4\delta\left(p_i(2|2) - \hat{p}_i(2)\right)
\end{aligned}
$$

Using $p_i(2|2) \geq p_i(2) + \lambda$ and $\hat{p}_i(2) \leq p_i(2) + \epsilon$, I derive lower bound

$$\Delta U_i(x_i = 2|S_i = 2) = 4\delta\left(p_i(2|2) - \hat{p}_i(2)\right) \geq 4\delta\left(p_i(2) + \lambda - (p_i(2) + \epsilon)\right) = 4\delta\left(\lambda - \epsilon\right)$$

on the gain in expected payment from reporting truthfully.

(Case 2): $|\hat{p}_i(2) - p_i(2)| > \epsilon$. (Also see Figure 6.2.)

In this case I derive an upper bound on the expected benefit from lying.

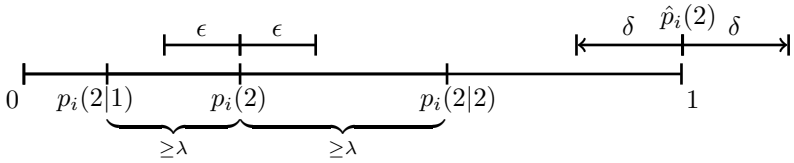


Figure 6.2: Illustration of Case 2 in the analysis of the Empirical Shadowing Method. Since $\hat{p}_i(2)$ is sampled using a finite number of samples and since it is not restricted to be within the ϵ bounds, it can happen that $\hat{p}_i(2) = 1$.

Consider again without loss of generality that $S_i = 2$:

$$\Delta U_i(x_i = 1|S_i = 2) = -\Delta U_i(x_i = 2|S_i = 2) = 4\delta\left(\hat{p}_i(2) - p_i(2|2)\right).$$

The maximal $\Delta U_i(x_i = 1|S_i = 2)$ is obtained for $\hat{p}_i(2)$ maximal and $p_i(2|2)$ minimal. Since $|\hat{p}_i(2) - p_i(2)| > \epsilon$, nothing prevents $\hat{p}_i(2) = 1$. From Lemma 6.3 we know that a lower bound of $p_i(2|2)$ given $p_i(2|2) - p_i(2) \geq \lambda$ and $p_i(2) - p_i(2|1) \geq \lambda$ is $2\sqrt{\lambda} - \lambda$, so that we can derive an upper bound for the expected benefit from lying by setting $\hat{p}_i(2) = 1$ and $p_i(2|2) = 2\sqrt{\lambda} - \lambda$, to obtain:

$$\Delta U_i(x_i = 1|S_i = 2) = 4\delta\left(\hat{p}_i(2) - p_i(2|2)\right) \leq 4\delta\left(1 - 2\sqrt{\lambda} + \lambda\right).$$

From Hoeffding's inequality, we know that Case 1 occurs with probability at least $1 - d$, so that the mechanism is truthful if

$$\begin{aligned}
&(1-d)4\delta(\lambda-\epsilon) > d4\delta\big(1-2\sqrt{\lambda}+\lambda\big)\\
\Leftrightarrow\ &(1-d)(\lambda-\epsilon) > d\big(1-2\sqrt{\lambda}+\lambda\big)\\
\Leftrightarrow\ &(\lambda-\epsilon) > d\,\underbrace{\big(1+2(\underbrace{\lambda-\sqrt{\lambda}}_{\geq -0.25}) - \underbrace{\epsilon}_{<\lambda\leq 0.5}\big)}_{>0}\\
\Leftrightarrow\ &d < \frac{\lambda-\epsilon}{1+2(\lambda-\sqrt{\lambda})-\epsilon}
\end{aligned}$$

To determine the number of items from which signals need to be sampled, the overall optimization problem becomes

$$\begin{aligned}
\text{min.}\quad & g\\
\text{s.t.}\quad & \epsilon < \lambda\\
& d < \tfrac{\lambda-\epsilon}{1+2(\lambda-\sqrt{\lambda})-\epsilon}\\
& g - 1 \geq \tfrac{1}{2\epsilon^2}\ln(\tfrac{2}{d}).
\end{aligned}$$

The last line contains $g - 1$ instead of g because $\hat{p}_i(2)$ was computed using

samples from $g-1$ items. For any fixed ϵ, it is optimal to maximize d, since this makes the right hand side of the final inequality as small as possible. Because of this, the problem can be restated as

$$\begin{aligned} \text{min.} \quad & g \\ \text{s.t.} \quad & \epsilon < \lambda \\ & d = \tfrac{\lambda-\epsilon}{1+2(\lambda-\sqrt{\lambda})-\epsilon} \\ & g-1 > \tfrac{1}{2\epsilon^2}\ln(\tfrac{2}{d}), \end{aligned}$$

where I have adopted equality for the second constraint and made the final inequality strict. Now, substituting for d in the last inequality, we obtain:

$$\begin{aligned} \text{min.} \quad & \tfrac{1}{2\epsilon^2}\ln\left(\tfrac{2\left(1+2(\lambda-\sqrt{\lambda})-\epsilon\right)}{\lambda-\epsilon}\right)+1 \\ \text{s.t.} \quad & \epsilon < \lambda \end{aligned}$$

This completes the proof. □

It is important to understand that the mechanism allows for subjective signal priors because it uses only "objective" signal reports which stem from the true item state to learn the signal prior. In particular, it does not elicit any beliefs from the agents. Since the signal reports used for learning the signal prior are not revealed to the agent, an agent forms a belief about this learned signal prior using her own subjective belief model and it is therefore sufficient that the derived bounds hold for any belief model that satisfies belief change bound λ.

Also note that for any given λ, the minimal number of required samples can be computed numerically. For example, given bound $\lambda = 0.05$, the optimal ϵ is $\epsilon = 0.046$, giving a corresponding requirement of $g-1 = 1351$ samples. I believe sample numbers in this range are reasonable for applications such as eliciting votes on the quality of an image label or whether a website is inappropriate for an advertiser. Note that these samples are from different items, so that this requires that there are many images or websites and not that there are many votes on any particular image or website.

6.5 Conclusion

In this chapter, I presented the first incentive compatible peer prediction mechanism that combines subjective belief models with minimal reporting. This combination is compelling because it provides robustness against strategic agents with non-standard (and possibly wrong) beliefs without requiring truthful agents

to deliberate about their beliefs. In the analysis of the Empirical Shadowing Method, I derive an upper bound on the number of items one needs to sample from. This mechanism could already be applied in applications such as crowdsourced image tagging, where requesters elicit information about many different items.

In future work, I plan to tighten the current analysis in regard to the number of samples required for strict incentives. The requirement that agent i has to expect that the learned signal prior $\hat{p}_i(2)$ is strictly in between the two possible posteriors is too pessimistic. Similar to the reasoning in the 2-Agent RBTS (Section 4.6) and the analysis of the Candidate shadow subjective-prior peer prediction mechanism (Section 5.4.2), the agent needs to build a weighted expected loss, i.e. compute the expected loss for each possible $\hat{p}_i(2)$ weighed with the probability of that $\hat{p}_i(2)$. Intuitively, if one instance of $\hat{p}_i(2)$ is much lower than $p_i(2)$ but another instance of $\hat{p}_i(2)$ is much higher than $p_i(2)$, then these may cancel each other out. Moreover, using this alternative approach, I expect to obtain analytical bounds that are stated just in terms of λ and not ϵ and λ.

Other interesting directions for future work are to extend the analysis to settings with multiple signals, and to design minimal and strictly truthful mechanisms for the orthogonal setting, where the mechanism has access to many signal reports from the *same* item (similar to the opinion poll setting introduced in Section 6.1).

Finally, it is interesting to continue to integrate machine learning models with peer prediction. One interesting area of application is peer grading in massively open online courses (MOOCs), where students grade other students' assignments. The machine learning work by Piech et al. [2013] learns each grader's quality and bias from Coursera[1] data with some success but ignores effort incentives for accurate grading. I believe that incorporating proper incentives for effort will increase the performance of these algorithms.

[1] `http://www.coursera.com`

Chapter 7

Effort Incentives with Fixed Costs

In this chapter, I study simple output-agreement mechanisms (slightly generalized from Section 2.4) in a setting different to the one studied in Chapters 2 to 6. The setting differs in three main aspects. First, instead of reporting on a single item, e.g. whether they would recommend a given hotel to a friend, agents are asked to report on the relative rank of two items, e.g. which of two hotels they consider the better choice. Second, it is assumed that agents differ in quality, which is defined as the probability that they can identify the correct, ground truth, ranking. This quality is at least 50%. It follows from this setup that simple output agreement is strictly truthful because after observing her signal, an agent's belief that her peer agent observed the same signal is larger than 50%. The third difference is that effort has a fixed cost, known to the mechanism. While all strictly truthful peer prediction mechanisms incentivize some cost of effort (Theorem 2.1), the particular cost that is incentivized depends on an agent's belief model and in particular how much the agent expects her belief to change because of a signal. The original peer prediction method (Chapter 2) can incorporate any fixed cost by scaling payments appropriately because the agents' belief model is common knowledge. The mechanisms from Chapters 3 to 6 relax the assumption that the mechanism knows the agents' belief models, and lose the ability to know how to appropriately scale payments.

The setting in this chapter is motivated by crowdsourcing human relevance judgments of search engine results, which are user studies run by search engine companies to test the performance of different ranking algorithms [e.g. Kazai et al., 2013]. In the simplest version of these relevance judgments, a user is presented with two websites together with a search query, and is asked to report

which of the two websites is better suited for the given query. These reports are then used to determine whether a change in the search engine's ranking algorithm would improve the quality of the search results and, if so, by how much. In another example, the two items are suggestions about things the New York City Mayor's Office could do to make New York a greener city.[1] In this wiki survey [Salganik and Levy, 2012], people were presented two different suggestions and were asked to vote which of the two they found more convincing. Note that depending on the choices presented to a user, she can be either very knowledgeable or clueless (or anything in between) as to which will best improve New York City's environment. The same holds true for human relevance judgments to improve search engine quality: depending on the query that is presented, a worker may or may not know how to evaluate a website based on that query.

It is important to point out the incentive problems in these settings. In the New York example, the truthfulness of reports may be a concern. A New York shop owner might agree that it makes the city cleaner charging customers 25 cents for each plastic bag they use, but she may also have a vested interest not to vote for this cause because it is bad for business. Another incentive issue is encouraging participants to invest costly effort in order to obtain information since people need to first understand the issue and then form an opinion about the presented options. In the human relevance judgments example, workers probably do not have strong incentives for being untruthful about which website they consider more appropriate for a given search term. It does, however, take effort clicking on the websites, having a closer look at each, and then weighing which of the two is better suited for a given query. Faced with a payment scheme that does not address this issue properly (such as a fixed-price scheme that pays a constant amount per reported ranking), workers would maximize their hourly wage not by investing effort but by randomly clicking through each task quickly.

The key property of an effort-incentivizing peer prediction mechanism is that the expected payment of an agent who invests effort is higher than the expected payment of an uninformed agent (an agent not investing effort) by at least the cost of effort. Without a lower bound on the extent that an agent's signal belief is expected to change following effort, there always exist stochastically relevant belief models and an effort cost $C > 0$ for which an agent will choose not to invest effort. Moreover, the problem cascades: when an uninformed agent is used as peer j for another agent i, this agent i no longer has an incentive to invest effort either.

I present two approaches to avoid this unraveling of incentives. To make progress, I first assume access to a quality oracle and the ability to block partic-

[1] `http://www.allourideas.org`

ipation based on an agent's quality. In the full model, I then allow an agent to first decide whether to "pass" and take a zero utility outside option, or participate. In addition, I allow for negative payments. Introducing the passing option and allowing for negative payments goes hand in hand because without negative payments every agent will choose to participate. The main result is that payments can be designed such that uninformed agents and agents whose qualities are too low are better off passing than participating while agents with qualities over a specific threshold that is a design parameter choose to invest effort and report their signals truthfully. That is, I design payments that incentivize agents to invest effort and self-select according to quality.

The remainder of this chapter is organized as follows. In Section 7.1, I discuss the difference to related work. In Section 7.2, I explain the difference of the model used in this chapter as compared to the models of Chapters 2 to 6. In Section 7.3, I analyze the basic decision problems in the model of this chapter from the perspective of a single agent investing effort and not investing effort. In Section 7.4, I develop a baseline mechanism through the assumption of access to a quality oracle that knows for every agent whether her quality is over a specified threshold and the mechanism lets only those agents over this threshold participate. I derive the mechanism's expected cost for, allowing for negative payments and restricting to non-negative payments. In Section 7.5, I then design a mechanism with negative payments that incentivizes agents to self-select, such that they only choose to participate when their quality is over the specified quality threshold. This self-selection mechanism has the same expected cost as the cost-minimal baseline mechanism with access to a quality oracle. In Section 7.6, I conclude with a discussion of the results and their implications for application, and point to interesting directions for future work.

7.1 Related Work

In addition to the related work on peer prediction mechanisms introduced in Chapters 2 to 6, there is additional related work particular to this chapter.

Most closely related is the work by Dasgupta and Ghosh [2013]. Their paper is situated in a very similar model for information elicitation with unknown ground truth where agents have individual qualities, defined just as in the model of this chapter as the probability that an agent can identify the correct, ground truth, ranking. The main difference to the model used in this chapter is that they require each agent to report on several items. While even simple output-agreement mechanisms induce a truthful equilibrium in their model, their key contribution is to develop a technique for eliminating other, unwanted equilibria. In a brief treatment, they comment that costly effort would be incentivized by

scaling payments and that the qualities could be obtained by pre-screening; e.g., through qualification tests. I formalize this approach to costly effort, and use the resulting mechanism as a baseline with which to compare my mechanism. The mechanism I introduce is shown to have significantly lower cost and does not require the ability to screen participants. Rather, participants will self-select into the mechanism according to their quality.

7.2 Model

There are two items A and B with true order $A \succ B$; a situation that is referred to as "A is best." Each agent in a sequence of agents is presented with the two items in a random order. From the perspective of a given agent i, the items are denoted A_i and B_i. For example, imagine that item A_i is the item presented on the left and B_i the item presented on the right, and that the decision about the ordering is made uniformly at random. Because of this, the signal prior of agent i is $\Pr(A_i \succ B_i) = 0.5$.

Agent i has the option of investing effort to observe a noisy signal $\succ_i$ about the true order of the items. In particular, agent i has a *quality* μ_i, which is drawn uniformly on $[0.5, 1]$. The distribution on quality is common knowledge, i.e. known to both the agents and the mechanism, but each agent's quality is private information of that respective agent. The cost of effort $C > 0$ is assumed to be identical for every agent, and common knowledge. If agent i invests effort, the signal she receives is the true order with probability μ_i; otherwise she receives the wrong order. For the analysis, it is convenient to transform quality $\mu_i \in [0.5, 1]$ to *normalized quality* $q_i \in [0, 1]$, so that $q_i = 2\mu_i - 1$ and q_i uniform on $[0, 1]$.

Each agent i is matched with another agent j, said to be the *peer* of agent i. For example, agent j can be the agent following agent i in the sequence, or agent j can be chosen randomly. Agent i can also be agent j's peer, but this need not be the case. I study simple output agreement mechanisms for which agent i receives payment $\tau_a > C > 0$ if her report agrees with that of peer agent j, and $\tau_d < \tau_a$ otherwise. This is a slight generalization from the definition of simple output agreement in Section 2.4 because it is not required that τ_d is set to 0. In fact, this chapter's focus is to consider the impact of allowing $\tau_d < 0$. To allow for negative payments in practice, one can imagine that the broader context requires holding at least $-\tau_d > 0$ as collateral from an agent who participates. The payments in the case of agreement and disagreement are common knowledge.

In contrast to the setting of Chapters 2 to 6, where every agent had an incentive to participate in the mechanism because all payments were non-negative, the

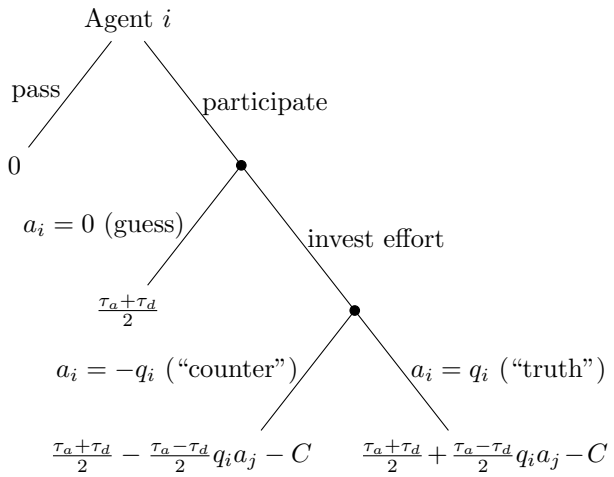


Figure 7.1: An illustration of agent i's decisions within the game.

participation step is modeled explicitly in this chapter. With negative payments, however, an agent may be better off obtaining zero utility by not participating. An agent first decides whether to "participate" or "pass," given knowledge of her quality μ_i. If she passes, she receives zero payment. If she participates, her strategic choice is whether to invest effort or not. The report of agent i is denoted by $\succ_i'$, and she must report $A_i \succ_i' B_i$ or $B_i \succ_i' A_i$. When saying that the report is $A \succ_i' B$ this should be understood to mean that the agent's report on A_i and B_i was mapped to mean $A \succ_i' B$. Only a participating agent can be chosen to be the peer of agent i, and if only one agent participates, her expected payment is assumed to be $(\tau_a + \tau_d)/2$, and the same as the payment she would obtain if she were matched against a peer who guesses.

7.3 Single-Agent Perspective

In this section, I analyze the set of possible best responses of an agent. We will need this in later sections when analyzing the equilibria of the mechanisms. See Figure 7.1 for a graphical illustration of the game from agent i's perspective.

7.3.1 Agent not Investing Effort

Consider an agent who chooses to participate but not invest effort. Because items A_i and B_i are in a random bijection to A and B, no matter if the agent reports $A_i \succ_i' B_i$, reports $B_i \succ_i' A_i$, or reports $A_i \succ_i' B_i$ with some probability, her belief about the effect of these reports is that it is equally likely to be $A \succ_i' B$ or $B \succ_i' A$. In the same way, an agent who does not invest effort will think it is

equally likely that peer agent j's report will correspond to $A_i \succ'_j B_i$ or $B_i \succ'_j A_i$ (with respect to agent i's item space). That is, the belief of agent i in regard to the report of her peer is $\Pr(A_i \succ'_j B_i) = 0.5$. For this reason, any uninformed reporting strategy comes down to guessing uniformly, with agent i receiving expected utility

$$U_i(\text{guess}) = \frac{\tau_a + \tau_d}{2} \tag{7.1}$$

for any report of her peer j.

Recall that the utility from not participating (i.e. from *passing*) is assumed to be zero:

$$U_i(\text{pass}) = 0. \tag{7.2}$$

Lemma 7.1 describes the primary effect of the mechanism parameters on the agents' equilibrium play:

Lemma 7.1. *Whether passing dominates guessing depends on payments τ_a and τ_d:*

1. *If $\tau_d > -\tau_a$, then $U_i(guess) > U_i(pass)$.*

2. *If $\tau_d < -\tau_a$, then $U_i(pass) > U_i(guess)$.*

Proof. $U_i(\text{guess}) = \frac{\tau_a+\tau_d}{2} > 0 = U_i(\text{pass}) \Leftrightarrow \tau_d > -\tau_a$ (second case analogous). □

For $\tau_d = -\tau_a$, agents are indifferent between passing and guessing. Whether passing dominates guessing or vice versa is one of the two major differences between the mechanisms of Sections 7.4 and 7.5.

7.3.2 Agent Investing Effort

First observe that Lemma 7.1 still holds after investing effort, but that investing effort followed by guessing or passing cannot be part of any equilibrium because effort is costly.

Having invested effort, an agent can now follow more informed reporting strategies. The *truth* strategy reports the true signal received. The *counter* strategy reports the opposite order to the one received. Many other reporting strategies are available. For example, agent i can report $A_i \succ'_i B_i$ with 50% probability if her signal is $A_i \succ_i B_i$ and report $B_i \succ'_i A_i$ otherwise. Lemma 7.2 says that these other reporting strategies following a decision to invest effort are not part of any equilibrium:

Lemma 7.2. *If investing effort is part of a best response for an agent, then the reporting strategies "truth" or "counter" strictly dominate all other strategies.*

Proof. Since $C > 0$, and given that investing effort is assumed part of a best response, the expected utility from investing effort must be higher than guessing. Therefore, the probability of agreement conditioned on at least one signal must be greater than 0.5. Before investing effort, the agent's subjective belief on j's report is $\Pr_i(A_i \succ'_j B_i) = 0.5$. Now suppose that this belief does not change after observing $A_i \succ_i B_i$, i.e. $\Pr_i(A_i \succ'_j B_i | A_i \succ_i B_i) = 0.5$. This would then mean that $\Pr_i(A_i \succ'_j B_i | B_i \succ_i A_i) = 0.5$ as well since

$$\begin{aligned} &\Pr_i(A_i \succ'_j B_i | A_i \succ_i B_i) \cdot \Pr_i(A_i \succ_i B_i) \\ +&\Pr_i(A_i \succ'_j B_i | B_i \succ_i A_i) \cdot \Pr_i(B_i \succ_i A_i) \\ =&\Pr_i(A_i \succ'_j B_i) = 0.5. \end{aligned} \tag{7.3}$$

But then agent i's subjective belief about the probability of agreement remains unchanged and we have a contradiction. Therefore, we must have $\Pr_i(A_i \succ'_j B_i | A_i \succ_i B_i) \neq 0.5$. Suppose $\Pr_i(A_i \succ'_j B_i | A_i \succ_i B_i) > 0.5$ and so $\Pr_i(A_i \succ'_j B_i | B_i \succ_i A_i) < 0.5$ (follows from Equation 7.3). Because of this, given signal $A_i \succ_i B_i$, the agent's unique best response is to report $A_i \succ'_i B_i$ and given signal $B_i \succ_i A_i$ her unique best response is to report $B_i \succ'_i A_i$. In each case, this "truth" strategy dominates any other strategy including a mixed strategy. Similarly, if $\Pr_i(A_i \succ'_j B_i | A_i \succ_i B_i) < 0.5$ and so $\Pr_i(A_i \succ'_j B_i | B_i \succ_i A_i) > 0.5$, then the "counter" strategy dominates any other strategy. □

Given Lemma 7.2, it is helpful to define the *action* a_i to succinctly represent all potential equilibrium play of an agent who chooses to participate. This action is defined as follows:

$$a_i = \begin{cases} q_i & \text{, if invest effort and report truth} \\ -q_i & \text{, if invest effort and report counter} \\ 0 & \text{, if no effort, and guess.} \end{cases} \tag{7.4}$$

Suppose, for example, that both agent i and her peer agent j invest effort and report truthfully. Agent i's expected utility, written as a function of a_i and

a_j, and when both invest effort and report truthfully is

$$\begin{aligned} U_i(a_i = q_i, a_j = q_j) = & \left(\frac{1+q_i}{2}\right)\left(\frac{1+q_j}{2}\right)\tau_a \\ & + \left(\frac{1-q_i}{2}\right)\left(\frac{1-q_j}{2}\right)\tau_a \\ & + \left(\frac{1+q_i}{2}\right)\left(\frac{1-q_j}{2}\right)\tau_d \\ & + \left(\frac{1-q_i}{2}\right)\left(\frac{1+q_j}{2}\right)\tau_d - C \\ = & \frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2} q_i q_j - C, \end{aligned} \tag{7.5}$$

where the first term is when both agents agree on the correct order, the second term is when both agree on the incorrect order, the third term is from agent i being correct but agent j being wrong, and the fourth line from agent i being wrong and agent j being correct.

On the other hand, if both invest effort but agent i plays truth and her peer plays counter, then τ_a and τ_d are simply exchanged, and:

$$U_i(a_i = q_i, a_j = -q_j) = \frac{\tau_a+\tau_d}{2} - \frac{(\tau_a-\tau_d)}{2} q_i q_j - C. \tag{7.6}$$

Suppose both were to invest effort and play counter. In this case, we again have:

$$U_i(a_i = -q_i, a_j = -q_j) = \frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2} q_i q_j - C. \tag{7.7}$$

Note also that if agent i invests effort, while her peer agent guesses, then her expected utility is just,

$$U_i(a_i = q_i, a_j = 0) = \frac{\tau_a+\tau_d}{2} - C. \tag{7.8}$$

Let σ_j denote peer agent j's strategy, which is a probability distribution over a_j. The expected valued of a_j given agent j's strategy is $\mathbf{E}[a_j|\sigma_j]$ .Combining Equations 7.5–7.8, we have the following lemma:

Lemma 7.3. *The expected utility for a participating agent with normalized quality q_i who participates and takes action $a_i \in \{-q_i, +q_i\}$ is:*

$$U_i(a_i) = \frac{\tau_a+\tau_d}{2} + \frac{\tau_a-\tau_d}{2} \cdot a_i \cdot \mathbf{E}[a_j|\sigma_j] - C, \tag{7.9}$$

where $\mathbf{E}[a_j|\sigma_j]$ is the expected value of the action of peer agent j and where $sigma_j$ denotes her strategy.

The expectation is taken with respect to the distribution on qualities of

agents, any mixing of strategies, and the random process that defines an agent's signal. Following effort by agent i, her strategic interaction with her peer agent j is precisely captured through $\mathbf{E}[a_j|\sigma_j]$, and Equation 7.9 fully captures agent i's expected utility following effort.

7.4 Quality-Oracle Mechanism

In this section, I analyze mechanisms with $\tau_d > -\tau_a$ and assume access to a quality oracle. In Section 7.5, we will then see that the right choice of $\tau_d < -\tau_a$ induces agents to self-select according to quality so that the mechanism no longer needs a quality oracle. The distinction between the cases $\tau_d > -\tau_a$ and $\tau_d < -\tau_a$ comes from Lemma 7.1. For $\tau_d > -\tau_a$, one needs some form of external quality screening because guessing strictly dominates passing, so that, without screening, every agent would participate, and low-quality agents would guess instead of investing effort and reporting truthfully. Knowing that some peer agents will guess, higher-quality agents would also guess, leading to an unraveling of incentives and additional noise in the reported signals.

Definition 15. For any qualification threshold $q^* > 0$, a mechanism with access to a *quality oracle* knows for every agent i whether normalized quality $q_i \geq q^*$ or $q_i < q^*$.

In the remainder of this section, it is assumed that the mechanism has access to a quality oracle, and uses this to only allow an agent to participate if her (normalized) quality is $q_i \geq q^* > 0$, for some threshold q^*.

When ground truth data is available for some item pairings, qualification tests can be used as an approximation for a quality oracle. Qualification tests ask every agent for reports about k item pairings for which the mechanism knows ground truth. Based on this, only those agents who agree with the ground truth on at least a fraction $q^* > 0$ of their reports are allowed to participate. Of course, such a qualification test provides only an approximate quality oracle. With $k = 10$ trials and a qualification threshold of $q^* = 0.6$ (corresponding to $\mu^* = 0.8$), for example, a qualification test allows an agent with $q_i = 0.2$ (corresponding to $\mu_i = 0.6$) to participate with 16.73% probability despite $q_i < q^*$. Similarly, an agent with $q_i = 0.7$ (corresponding to $\mu_i = 0.85$) misses the qualification bar with 17.98% probability. Due to the law of large numbers, these mis-classifications disappear for $k \to \infty$, so that a qualification test with $k \to \infty$ trials approaches the behavior of a quality oracle.

7.4.1 Incentive Analysis

Theorem 7.4. *With $\tau_d > -\tau_a$ and $\tau_a - \tau_d > \frac{4C}{(q^*)^2+q^*}$, the mechanism with access to a quality oracle induces a strict Bayes-Nash equilibrium where every agent i allowed to participate, i.e. with quality $q_i \geq q^*$, chooses to participate, invest effort, and report truthfully.*

Proof. Consider an agent i who is allowed to participate after the mechanism asked the quality oracle, and assume all other agents who are allowed to participate invest effort and report truthfully. To prove is that agent i's unique best response is to also invest effort and report truthfully. Using Equation 7.9, we first need to compute $\mathbf{E}[a_j|\sigma_j]$, where agent j invests effort and reports truthfully:

$$\mathbf{E}[a_j|\sigma_j] = \mathbf{E}[q_j|q_j \geq q^*] = \frac{1}{1-q^*}\int_{q_j=q^*}^{1} q_j \, \mathrm{d}q_j = \frac{1+q^*}{2}$$

Observe that the quality distribution for peer agent j is now uniform on $[q^*, 1]$.

Since $\tau_d > -\tau_a$, we know from Lemma 7.1 that passing is strictly dominated by guessing, so that in every equilibrium agent i is always participating. From Lemma 7.2 we know that only $a_i \in \{-q_i, 0, +q_i\}$ can be best responses of an agent that participates. Now since $\mathbf{E}[a_j|\sigma_j] = \frac{1+q^*}{2} > 0$ and $q_i \geq q^* > 0$, by Equation 7.9, $a_i = -q_i$ cannot be part of a best response either. It then remains to determine the values τ_a, τ_d with $\tau_d > -\tau_a$ for which an agent i with quality $q_i = q^*$ is better off playing $a_i = q_i$ than $a_i = 0$ by setting $U_i(a_i = q^*) > U_i(a_i = 0)$:

$$\begin{aligned} &\frac{\tau_a + \tau_d}{2} + \frac{\tau_a - \tau_d}{2} \cdot q^* \cdot \mathbf{E}[a_j|\sigma_j] - C > \frac{\tau_a + \tau_d}{2} \\ \Leftrightarrow \frac{\tau_a - \tau_d}{2} \cdot \frac{q^*(1+q^*)}{2} > C &\Leftrightarrow \tau_a - \tau_d > \frac{4C}{(q^*)^2 + q^*}. \end{aligned}$$

This completes the proof. □

7.4.2 Expected Cost

Now that we know the constraint on τ_a and τ_d such that for a given quality threshold q^*, there is a Bayes-Nash equilibrium where all agents with quality higher than the threshold invest effort and are truthful, how should payments τ_a and τ_d be set to minimize the expected cost given C and q^*?

For each agent who participates, the expected cost in the truthful equilibrium

of the quality-oracle mechanism is given by:

$$\begin{aligned}
&\mathbf{E}[\text{cost for participating agent}] \\
=&\frac{\tau_a+\tau_d}{2}+\frac{\tau_a-\tau_d}{2}\cdot\frac{q^*+1}{2}\cdot\frac{q^*+1}{2} \\
=&\frac{\tau_a+\tau_d}{2}+\frac{\tau_a-\tau_d}{2}\cdot\frac{(q^*+1)^2}{4} \\
=&\frac{1}{2}(\tau_a+\tau_d)+\frac{(q^*+1)^2}{8}(\tau_a-\tau_d).
\end{aligned} \tag{7.10}$$

Given this, the optimization problem to find the cost-minimizing mechanism parameters becomes:

$$\begin{aligned}
\text{minimize}\quad & \frac{1}{2}(\tau_a+\tau_d)+\frac{(q^*+1)^2}{8}(\tau_a-\tau_d) \\
\text{s.t.}\quad & \tau_a-\tau_d>\frac{4C}{(q^*)^2+q^*} \\
& \tau_a+\tau_d>0
\end{aligned} \tag{7.11}$$

The remainder of the section is organized as follows. I first solve this optimization problem, and then impose the additional requirement of non-negative payments, i.e. $\tau_d \geq 0$. Having done this, I quantify how much more the mechanism has to pay in expectation because of this restriction.

Allowing for Negative Payments

I solve the optimization problem as given in (7.11) using a variable change. Let $\tau_a = \tau + \Delta$ and $\tau_d = \tau - \Delta$ for new variables τ and Δ. Substituting into (7.11) and solving the optimization problem immediately gives $\Delta = 2C/\big((q^*)^2+q^*\big)+\epsilon$, and $\tau = \epsilon$, with $\epsilon > 0$ and $\epsilon \to 0$. Substituting again for τ_a and τ_d, we obtain the cost-minimizing mechanism parameters:

$$\tau_a = \frac{2C}{(q^*)^2+q^*}+\epsilon \tag{7.12}$$

and

$$\tau_d = -\frac{2C}{(q^*)^2+q^*}. \tag{7.13}$$

Given this, the expected cost to the mechanism for each agent who chooses

to participate is

$$\begin{aligned}&\mathbf{E}[\text{minimal cost for participating agent}|\tau_d > -\tau_a]\\ =&\frac{1}{2}(\tau_a+\tau_d)+\frac{(q^*+1)^2}{8}(\tau_a-\tau_d)\\ =&\frac{1}{2}(\epsilon)+\frac{(q^*+1)^2}{8}\left(\frac{4C}{(q^*)^2+q^*}+\epsilon\right),\end{aligned} \tag{7.14}$$

and for $\epsilon \to 0$, we have:

$$\begin{aligned}&\lim_{\epsilon\to 0}\mathbf{E}[\text{minimal cost for participating agent}|\tau_d > -\tau_a]\\ =&\frac{(q^*+1)^2}{8}\cdot\frac{4C}{(q^*)^2+q^*}=\frac{(q^*)^2+2q^*+1}{2(q^*)^2+2q^*}\;C\\ =&\left(\frac{1}{2q^*}+\frac{1}{2}\right)\;C.\end{aligned} \tag{7.15}$$

Requiring Non-negative Payments

Let us now suppose that we seek to minimize the expected cost subject to $\tau_d \geq 0$. The second constraint in (7.11) is always satisfied with $\tau_d \geq 0$. I now argue that it must be that $\tau_d = 0$ in an optimal solution. Assume this was not the case, so that $\tau_d = z$ for some $z > 0$. Then the left-hand side of the first constraint can be kept at the same level by setting $\tau_d := 0$ and $\tau_a := \tau_a - z$, which would lower the first part of the objective function and leave the second part unchanged. So for minimal non-negative payments, we have:

$$\tau_d = 0. \tag{7.16}$$

After inserting $\tau_d = 0$ back into the optimization problem, we obtain:

$$\tau_a = \frac{4C}{(q^*)^2+q^*}+\epsilon \tag{7.17}$$

with $\epsilon \to 0$. Based on this, the expected cost to the mechanism for each agent who chooses to participate is

$$\begin{aligned}&\mathbf{E}[\text{minimal cost for participating agent}|\tau_d \geq 0]\\ =&\frac{1}{2}(\tau_a+\tau_d)+\frac{(q^*+1)^2}{8}(\tau_a-\tau_d)\\ =&\frac{(q^*)^2+2q^*+5}{8}\left(\frac{4C}{(q^*)^2+q^*}+\epsilon\right)\end{aligned} \tag{7.18}$$

and for $\epsilon \to 0$, we have:

$$
\begin{aligned}
&\lim_{\epsilon \to 0} \mathbf{E}[\text{minimal cost for participating agent} | \tau_d \geq 0] \\
=&\frac{(q^*)^2 + 2q^* + 5}{8} \left(\frac{4C}{(q^*)^2 + q^*} + \epsilon \right) \\
=&\frac{(q^*)^2 + 2q^* + 5}{2(q^*)^2 + 2q^*} C.
\end{aligned} \tag{7.19}
$$

Relative Cost of Requiring Non-Negative Payments

Constraining the mechanism's payments to be non-negative can only increase the expected cost (fixing the cost to agents for effort C and the design parameter q^* above which agents will choose to participate and invest effort). But how much more expensive is it when the mechanism is restricted in this way? Recall that in both cases agents must have an incentive to participate, i.e. the expected utility for an agent who participates must be non-negative.

Theorem 7.5. *Fixing quality threshold q^*, the expected cost of the cost-optimized quality-oracle mechanism increases by a factor of*

$$\frac{4}{(q^* + 1)^2} + 1$$

when constraining the mechanism to non-negative payments $\tau_a, \tau_d \geq 0$. This is an increase between 2 (for $q^* \to 1$) and 5 (for $q^* \to 0$).

Proof. The result follows from dividing Equation 7.19 by Equation 7.15:

$$
\begin{aligned}
&\frac{\lim_{\epsilon \to 0} \mathbf{E}[\text{minimal cost for participating agent} | \tau_d \geq 0]}{\lim_{\epsilon \to 0} \mathbf{E}[\text{minimal cost for participating agent} | \tau_d > -\tau_a]} \\
=&\frac{\frac{(q^*)^2 + 2q^* + 5}{2(q^*)^2 + 2q^*} C}{\frac{(q^*)^2 + 2q^* + 1}{2(q^*)^2 + 2q^*} C} = \frac{(q^*)^2 + 2q^* + 5}{(q^*)^2 + 2q^* + 1} \\
=&\frac{(q^* + 1)^2 + 4}{(q^* + 1)^2} = \frac{4}{(q^* + 1)^2} + 1
\end{aligned} \tag{7.20}
$$

Since $(q^* + 1)^2$ is strictly increasing in q^*, the term $4/(q^* + 1)^2 + 1$ is strictly decreasing. The statement follows after inserting $q^* = 0$ and $q^* = 1$. □

7.5 Self-Selection Mechanism

In this section, I drop the assumption that the mechanism has access to a quality oracle. At the same time, I consider the effect of setting $\tau_d < -\tau_a$ so that passing dominates guessing (Lemma 7.1). The main result is the identification of an equilibrium in which agents self-select according to their quality q_i, such that

every agent over quality threshold q^* invests effort and is truthful, and every agent below the threshold is passing. While optimizing the quality threshold is left for future work, I note here that the quality threshold trades off quality, cost and number of reports: a higher threshold q^* results in higher expected quality of reports and lower expected cost per report but also results in more agents not participating and thus fewer reports.

There are several advantages of self selection when compared to qualification tests. First, qualification tests are equivalent to quality oracles only with infinitely many samples. Second, qualification tests are wasteful because agents need to be paid for test questions to which the answer is already known. Third, self selection is more flexible than qualification tests in that it adapts to changes in the nature of tasks without any re-testing. In the human relevance judgments setting, for example, the quality of an agent does not have to be fixed but can depend on the given search query. Finally, self selection does not require ground truth data.

7.5.1 Incentive Analysis

Let $\sigma_i(q_i)$ denote the strategy that maps agent i's quality type q_i to her action (e.g. "pass" or "guess").

Theorem 7.6. *With $\tau_d < -\tau_a$, the mechanism without access to a quality oracle induces a Bayes-Nash equilibrium where every agent i plays the following strategy:*

$$\sigma_i(q_i) = \begin{cases} \textit{invest effort and report truthfully}, & \textit{if } q_i \geq q^* \\ \textit{pass}, & \textit{if } q_i < q^* \end{cases}$$

where

$$q^* = \sqrt{\frac{9}{4} - \frac{4(\tau_a - C)}{\tau_a - \tau_d}} - \frac{1}{2}.$$

For an agent with $q_i \neq q^*$ this is a strict best response.

Proof. It is sufficient to show that given peer agent j plays $\sigma_j(q_j)$, it is a best response for agent i to play $\sigma_i(q_i)$, and the unique best response if $q_i \neq q^*$. Inserting $\sigma_j(q_j)$ into Equation 7.9, $\mathbf{E}[a_j|\sigma_j]$ is identical to its value in the quality-oracle mechanism:

$$\mathbf{E}[a_j|\sigma_j] = \mathbf{E}[q_j|q_j \geq q^*] = \frac{1}{1-q^*} \int_{q_j=q^*}^{1} q_j \, \mathrm{d}q_j = \frac{1+q^*}{2}$$

By Lemma 7.1, we know that passing strictly dominates guessing, so in order to determine the indifference point between passing and investing effort followed

by truthful reporting, we set $U_i(a_i = q^*) = U_i(\text{pass}) = 0$ and obtain

$$\begin{aligned} & \frac{\tau_a + \tau_d}{2} + \frac{\tau_a - \tau_d}{2} \cdot q^* \cdot \mathbf{E}[a_j|\sigma_j] - C = 0 \\ \Leftrightarrow & \frac{\tau_a + \tau_d}{2} + \frac{(\tau_a - \tau_d)(q^*(1+q^*))}{4} - C = 0 \\ & \Leftrightarrow q^* + (q^*)^2 = 2 - \frac{4(\tau_a - C)}{\tau_a - \tau_d} \\ & \Leftrightarrow q^* = \pm\sqrt{\frac{9}{4} - \frac{4(\tau_a - C)}{\tau_a - \tau_d}} - \frac{1}{2}. \end{aligned} \tag{7.21}$$

First observe that $\frac{4(\tau_a - C)}{\tau_a - \tau_d} > 0$ because $\tau_a > C$. Now, $\tau_a - C < \tau_a$ and $\tau_a - \tau_d > 2\tau_a$ since $-\tau_d > \tau_a$. Therefore, it holds that $4(\tau_a - C)/(\tau_a - \tau_d) < 4\tau_a/2\tau_a = 2$ and so only the positive solution of the square root is within the $[0, 1]$ bounds for q^*. Strictness for $q_i \neq q^*$ follows from Equation 7.9 strictly increasing with a_i for $\mathbf{E}[a_j|\sigma_j] > 0$. □

7.5.2 Expected Cost

Since agent j's equilibrium play is the same as in the previous section, the equation for the expected cost of the mechanism is unchanged (Equation 7.10). However, the equilibrium conditions of the self-selection mechanism do not allow the same analysis used to find the cost-minimal payments in the quality-oracle mechanism. This is because the equilibrium condition from Theorem 7.6 does not have the same simple structure as that in Theorem 7.4. We thus insert the equilibrium condition into Equation 7.10 and obtain (the second line in Equation 7.22 is derived from the second line in Equation 7.21):

$$\begin{aligned} & \mathbf{E}[\text{cost for participating agent}] \\ =& \frac{\tau_a + \tau_d}{2} + \frac{\tau_a - \tau_d}{2} \cdot \frac{(q^* + 1)^2}{4} \\ =& \frac{\tau_a + \tau_d}{2} + \frac{2C - (\tau_a + \tau_d)}{q^* + (q^*)^2} \cdot \frac{(q^* + 1)^2}{4} \\ =& \frac{\tau_a + \tau_d}{2} + \frac{(q^* + 1)\big(2C - (\tau_a + \tau_d)\big)}{4q^*} \\ =& \frac{\tau_a + \tau_d}{2} - \frac{(\tau_a + \tau_d)(q^* + 1)}{4q^*} + \frac{2C(q^* + 1)}{4q^*} \\ =& \frac{\tau_a + \tau_d}{2} - \frac{\tau_a + \tau_d}{2}\left(\frac{1}{2} + \frac{1}{2q^*}\right) + \frac{C}{2} + \frac{C}{2q^*} - C + C \\ =& \frac{\tau_a + \tau_d}{2}\left(\frac{1}{2} - \frac{1}{2q^*}\right) - C\left(\frac{1}{2} - \frac{1}{2q^*}\right) + C \\ =& \left(\frac{\tau_a + \tau_d}{2} - C\right)\left(\frac{1}{2} - \frac{1}{2q^*}\right) + C = \left(C - \frac{\tau_a + \tau_d}{2}\right)\left(\frac{1}{2q^*} - \frac{1}{2}\right) + C \end{aligned} \tag{7.22}$$

Both factors of the first part of this equation are always positive for $\tau_d < -\tau_a$ and $q^* \in (0, 1)$. Fixing q^*, the minimal payments are thus setting $\tau_d = -\tau_a - \epsilon$ with $\epsilon > 0$ and $\epsilon \to 0$, so that $(\tau_a + \tau_d)/2 \to 0$. Setting $\tau_d = -\tau_a - \epsilon$ leaves us with one degree of freedom, and τ_a can still be used to implement any q^*, since

$$\lim_{\epsilon \to 0} \frac{4(\tau_a - C)}{\tau_a - \tau_d} = \lim_{\epsilon \to 0} \frac{4(\tau_a - C)}{2\tau_a + \epsilon} = 2 - \frac{2C}{\tau_a},$$

so that

$$q^* := \sqrt{\frac{9}{4} - \left(2 - \frac{2C}{\tau_a}\right)} - \frac{1}{2} = \sqrt{\frac{1}{4} + \frac{2C}{\tau_a}} - \frac{1}{2}, \tag{7.23}$$

can be set to any value between 0 and 1. Solving Equation 7.23 for τ_a, we thus obtain the cost-minimal payments

$$\tau_a = \frac{2C}{(q^*)^2 + q^*}$$

and

$$\tau_d = -\frac{2C}{(q^*)^2 + q^*} - \epsilon.$$

For $\epsilon \to 0$, these are identical to the cost-minimal payments of the quality-oracle mechanism with $\tau_d > -\tau_a$. Therefore, the self-selection mechanism with $\tau_d < -\tau_a$ has the same expected cost, with the added benefit of obtaining perfect screening through self-selection. Theorem 7.7 then follows immediately:

Theorem 7.7. *For fixed quality threshold q^*, the expected cost of the cost-optimized self-selection mechanism is lower than the expected cost of the cost-optimized quality-oracle mechanism constrained to non-negative payments by a factor of*

$$\frac{4}{(q^* + 1)^2} + 1.$$

7.6 Conclusion

In this chapter, I presented an analysis of simple output-agreement mechanisms for incentivizing effort and providing screening for worker quality. The analysis suggests that practitioners should strongly consider allowing negative payments as they significantly lower the cost for effort-incentivizing peer prediction mechanisms and provide a free way to perfectly screen based on quality in equilibrium.

In closing, I emphasize the two main contributions of this chapter and discuss the practicality of the presented approach. First, peer prediction with effort incentives is expensive if simple output agreement can only use non-negative payments. For example, with effort cost $C > 0$ and a quality threshold of $\mu^* = 0.8$ (i.e., $q^* = 0.6$ and thus screening for the top 40% of quality in the

market), the expected cost for the cost-minimal, non-negative-payment output-agreement mechanism is $3.4C$. Allowing for negative payments, the expected cost for the mechanism decreases to $1.3C$. This improvement would occur even if the designer had access to a free method of perfectly screening for the quality of participants and simply stems from using lower payments.

Second, in addition to lower expected cost on behalf of the mechanism, choosing negative payments to disincentivize guessing induces an equilibrium where agents self-select according to the selection criterion—in effect, perfect screening comes for free. I do not believe that this principle is restricted to selection for quality. For example, it could also be applied when all participants have the same quality but differ in cost, and participants self-select according to cost.

In markets where the requester competes with other requesters, a worker's outside option may not be to pass and obtain utility zero but to work on another task with positive expected utility. For such a competitive setting, I conjecture that the relative cost benefit of negative payments decreases but that incentivizing workers with low quality to pass still has the benefit that it induces workers to self select according to their qualities.

Regarding the practicality of the approach, it seems useful to separate the assumptions made for the analysis (such as the uniform prior on the agents' quality) and the simplicity (and thus practical robustness) of simple output-agreement mechanisms. In practice, a designer only needs to set two parameters, agreement payment τ_a and disagreement payment τ_d, and this could be achieved adaptively. The main theoretical results suggest the opportunity for significant cost savings, along with screening of agents according to quality through self selection.

Future Work

There are several interesting directions for future work. First, it would be interesting to evaluate the mechanism experimentally. Second, it would be interesting to extend the analysis to the more sophisticated belief models of Chapters 2 to 6, where an agent may believe that she holds a minority opinion after investing effort. This is currently precluded because an agent observes ground truth with probability larger than 50% after investing effort. I intend to study effort incentives for peer-prediction mechanisms that are not just simple output agreement, such as the Robust Bayesian Truth Serum (Chapter 4).

Chapter 8

Conclusion

Peer prediction mechanisms have originally been motivated with the need to truthfully elicit ratings about products or services in reputation systems, such as those employed by Yelp!,[1] Amazon, or eBay. In hindsight, this motivation seems problematic for several reasons. First, peer prediction mechanisms rely on payments, which are infeasible in many product rating environments. Second, the peer prediction approach does not seem appropriate for the diverse rater population of real-world reputation systems. The computed payments are either too high, in that raters are intrinsically motivated to report truthfully, even without payment, or they are too low in that the incentives to manipulate ratings are too large to be compensated through realistically scaled payments. As an example for this latter case, consider a restaurant owner who could increase her restaurant's rating on Yelp! by falsely reporting a positive experience. Applying a peer prediction mechanism to disincentivize this manipulation would require an infeasibly high scaling of peer prediction payments given the high impact of restaurant reviews on the restaurant's revenue [Luca, 2011].

Nevertheless, peer prediction mechanisms do address important incentive problems in other applications. In particular, they are useful in paid crowdsourcing settings, where participants need to invest costly effort. In Section 2.2, I presented an example for such a setting, in which crowd workers were asked to report whether a given website contains offensive content, such as explicit violence or nudity. Paid crowdsourcing is a natural application for peer prediction mechanisms because a payment infrastructure is already established. Moreover, while there are no incentives for a worker to lie, workers do have an incentive to shirk, and not invest effort, unless they are properly incentivized. As I have shown in this thesis, peer prediction mechanisms address this problem and provide the appropriate incentives for the investment of costly effort.

[1] `http://www.yelp.com`

There are many interesting directions for future work. The first is to extend the fixed-cost effort analysis from Chapter 7 to the more sophisticated belief models I have discussed in Chapters 2 to 6, where an agent may believe that she holds a minority opinion after investing effort.

Second, every truthful peer prediction mechanism also has non-truthful equilibria [Waggoner and Chen, 2013]. An important direction for future work is thus the design of a general technique to prevent agents from coordinating on any of these lying equilibria, which may also Pareto-dominate the truthful equilibrium [Jurca and Faltings, 2009]. Researchers have made first steps in this direction but the results are restricted. They either apply only to the common belief model of the classical peer prediction method [Jurca and Faltings, 2009] or only to simple output agreement settings [Dasgupta and Ghosh, 2013]. In particular, no solution is known for any of the general robust peer prediction models of Chapters 3 to 6. It will be interesting to see whether the early results can be generalized.

Third, in the models of Chapters 2 to 6, agents are assumed to be identical, in that the signals they observe only depend on the true world state. In reality, agents are likely to also have individual characteristics that influence their signals. A first step to incorporate such characteristics is the work presented in Chapter 7, where agents are allowed to be of different quality. An interesting direction for future work is to incorporate other realistic agent characteristics, such as taste differences and biases, into robust peer prediction. While there is work in the machine learning literature that incorporates these types of characteristics [e.g. Piech et al., 2013], this line of research ignores incentives. I believe that integrating peer prediction mechanisms with machine learning models will be mutually beneficial. It will be important to balance richer agent models with reporting costs that are still feasible for participants.

Fourth, in addition to these extensions to the standard peer prediction setting, another interesting direction for future work is to combine the presented robust peer prediction mechanisms with mechanisms used in other multiagent settings and classical mechanism design. An early example in this spirit is the work by Azar et al. [2012], who propose an auction where one of the participants is asked to predict the bids of others, and where this prediction is scored using a proper scoring rule. Using the predicted bids, a reserve price is computed, and this robust auction is then shown to approximate the optimal, revenue-maximizing auction that has knowledge of the bid distribution. A natural next step in combining robust peer prediction with classical mechanism design are mechanisms with correlated types, so that the private preferences of an agent tell her something about the private preferences of others.

Bibliography

Arrow, K. J. (1963). *Social Choice and Individual Values*. John Wiley & Sons, Inc., 2 edition.

Azar, P., Chen, J., and Micali, S. (2012). Crowdsourced Bayesian Auctions. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS'12)*, pages 236–248.

Battigali, P., Gilli, M., and Molinari, M. C. (1992). Learning convergence to equilibrium in repeated strategic interactions: An introductory survey,. *Ricerche Economiche*, 96:335–378.

Bergemann, D. and Morris, S. (2005). Robust Mechanism Design. *Econometrica*, 73(6):1771–1813.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3.

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

Crémer, J. and McLean, R. P. (1985). Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist when Demands are Interdependent. *Econometrica*, 53(2):345–361.

Crémer, J. and McLean, R. P. (1988). Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions. *Econometrica*, 56(6):1247–1257.

Dasgupta, A. and Ghosh, A. (2013). Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *Proceedings of the 22nd ACM International World Wide Web Conference (WWW'13)*, pages 319–330.

Domke, J. (2010). Statistical Machine Learning. `http://users.cecs.anu.edu.au/~jdomke/courses/sml2010/10theory.pdf`. Lecture Notes, Learning Theory (10).

Friedman, D. (1983). Effective Scoring Rules for Probabilistic Forecasts. *Management Science*, 29(4):447–454.

Gibbard, A. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4):587–602.

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

Goldstein, D. G., McAfee, R. P., and Suri, S. (2012). Improving the Effectiveness fo Time-Based Display Advertising. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*, pages 639–654.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B*, 14(1):107–114.

Green, J. R. and Laffont, J.-J. (1977). Characterization of Satisfactoy Mechanisms for the Revelation of Preferences for Public Goods. *Econometrica*, 45(2):427–438.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30.

Johnson, S., Pratt, J. W., and Zeckhauser, R. J. (1990). Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case. *Econometrica*, 58(4):873–900.

Jurca, R. and Faltings, B. (2005). Enforcing Truthful Strategies in Incentive Compatible Reputation Mechanisms. In *Proceedings of the 1st International Workshop on Internet and Network Economics (WINE'05)*, pages 268–277.

Jurca, R. and Faltings, B. (2006). Minimum Payments that Reward Honest Reputation Feedback. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06)*, pages 190–199.

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

Jurca, R. and Faltings, B. (2008). Incentives for Expressing Opinions in Online Polls. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08)*, pages 119–128.

Jurca, R. and Faltings, B. (2009). Mechanisms for Making Crowds Truthful. *Journal of Artificial Intelligence Research (JAIR)*, 34:209–253.

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

Kalai, E. and Lehrer, E. (1993). Subjective equilibrium in repeated games. *Econometrica*, 61:1231–1240.

Kazai, G., Kamps, J., and Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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

Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.com. Technical Report 12-016, Harvard Business School.

Matheson, J. E. and Winkler, R. L. (1976). Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10):1087–1096.

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

Myerson, R. B. (1979). Incentive Compatibility and the Bargaining Problem. *Econometrica*, 47(1):61–74.

Myerson, R. B. (1981). Optimal Auction Design. *Mathematics of Operations Research*, 6(1):58–73.

Nisan, N. (2007). Introduction to Mechanism Design (for Computer Scientists). In Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., editors, *Algorithmic Game Theory*, chapter 9, pages 209–242. Cambridge University Press.

Pennock, D. and Sami, R. (2007). Computational Aspects of Prediction Markets. In Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., editors, *Algorithmic Game Theory*, chapter 26, pages 651–676. Cambridge University Press.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM'13)*, pages 153–160. International Educational Data Mining Society.

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 *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13)*, pages 833–839.

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

Rubinstein, A. and Wolinsky, A. (1984). Rationalizable conjectural equilibrium: Between Nash and Rationalizability. *Games and Economic Behavior*, 6:299–311.

Salganik, M. J. and Levy, K. E. C. (2012). Wiki surveys: Open and quantifiable social data collection. preprint, `http://arxiv.org/abs/1202.0500`.

Satterthwaite, M. A. (1975). Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, 10:187–217.

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, 66:783–801.

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

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

Waggoner, B. and Chen, Y. (2013). Information Elicitation Sans Verification. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC'13)*.

Wilson, R. (1987). Game-Theoretic Analyses of Trading Processes. In Bewley, T. F., editor, *Advances in Economic Theory: Fifth World Congress*, chapter 2. Cambridge University Press.

Witkowski, J. (2009). Eliciting Honest Reputation Feedback in a Markov Setting. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09)*, pages 330–335.

Witkowski, J. (2010). Truthful Feedback for Sanctioning Reputation Mechanisms. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI'10)*, pages 658–665.

Witkowski, J. (2011a). Incentive-Compatible Trust Mechanisms. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, pages 1865–1866.

Witkowski, J. (2011b). Trust Mechanisms for Online Systems. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, pages 2866–2867.

Witkowski, J., Bachrach, Y., Key, P., and Parkes, D. C. (2013). Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*, pages 190–197.

Witkowski, J. and Parkes, D. C. (2012a). A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, pages 1492–1498.

Witkowski, J. and Parkes, D. C. (2012b). Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*, pages 964–981.

Witkowski, J. and Parkes, D. C. (2013). Learning the Prior in Minimal Peer Prediction. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC'13)*.

Witkowski, J., Seuken, S., and Parkes, D. C. (2011). Incentive-Compatible Escrow Mechanisms. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, pages 751–757.

Wolfers, J. and Zitzewitz, E. (2004). Prediction Markets. *Journal of Economic Perspectives*, 18(2):107–126.