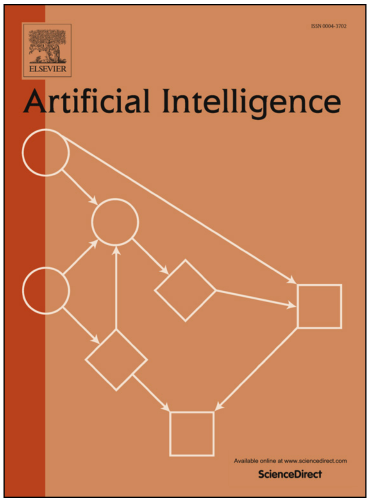

Incentivizing Evaluation with Peer Prediction and Limited Access to Ground Truth

Xi Alice Gao, James R. Wright, Kevin Leyton-Brown

PII: S0004-3702(19)30155-9
DOI: https://doi.org/10.1016/j.artint.2019.03.004
Reference: ARTINT 3153

To appear in: *Artificial Intelligence*

Received date: 12 September 2016
Revised date: 5 March 2019
Accepted date: 14 March 2019

Please cite this article in press as: X.A. Gao et al., Incentivizing Evaluation with Peer Prediction and Limited Access to Ground Truth, *Artif. Intell.* (2019), https://doi.org/10.1016/j.artint.2019.03.004

Incentivizing Evaluation with Peer Prediction and Limited Access to Ground Truth

Xi Alice Gao

University of Waterloo, Waterloo, Ontario, Canada

James R. Wright

University of Alberta, Edmonton, Alberta, Canada

Kevin Leyton-Brown

University of British Columbia, Vancouver, British Columbia, Canada

Abstract

In many settings, an effective way of evaluating objects of interest is to collect evaluations from dispersed individuals and to aggregate these evaluations together. Some examples are categorizing online content and evaluating student assignments via peer grading. For this data science problem, one challenge is to motivate participants to conduct such evaluations carefully and to report them honestly, particularly when doing so is costly. Existing approaches, notably peer-prediction mechanisms, can incentivize truth telling in equilibrium. However, they also give rise to equilibria in which agents do not pay the costs required to evaluate accurately, and hence fail to elicit useful information. We show that this problem is unavoidable whenever agents are able to coordinate using low-cost signals about the items being evaluated (e.g., text labels or pictures). We then consider ways of circumventing this problem by comparing agents' reports to ground truth, which is available in practice when there exist trusted evaluators—such as teaching assistants in the peer grading scenario—who can perform a limited number of unbiased (but noisy) evaluations. Of course, when such ground truth is available, a simpler approach is also possible: rewarding each agent based on agreement with ground truth with some probability, and unconditionally rewarding the agent otherwise. Surprisingly, we show that the simpler mechanism achieves stronger incentive guarantees given less access to ground truth than a large set of peer-prediction mechanisms.

Keywords: peer prediction, peer grading, incentivize effort, incentivize truthful reporting, information elicitation, game theory, crowdsourcing.

1. Introduction

In many practical settings, an effective way of evaluating objects of interest is to collect evaluations from a large, dispersed group of agents. These evaluations can then be aggregated together and provided as a service, as in online forums such as Rotten Tomatoes, Yelp, and TripAdvisor, which obtain feedback from millions of users about movies, restaurants, and travel destinations. Alternatively, the aggregated evaluations can be used directly. For example, crowdsourcing platforms are increasingly used to collect semantic labels of images and online content for use in training machine learning algorithms.

However, these agents may not be motivated to invest costly effort to obtain accurate evaluations. Therefore, an important problem in artificial intelligence is to design incentives to motivate large groups of agents to obtain and to reveal accurate information (e.g., Prelec, 2004; Miller et al., 2005; Zohar & Rosenschein, 2006, 2008; Jurca & Faltings, 2009; Papakonstantinou et al., 2008, 2010, 2011; Faltings et al., 2012; Witkowski & Parkes, 2012; Witkowski et al., 2013; Dasgupta & Ghosh, 2013; Witkowski & Parkes, 2013; Radanovic & Faltings, 2013; Shah et al., 2013; Radanovic & Faltings, 2014; Radanovic et al., 2016; Riley, 2014; Zhang & Chen, 2014; Waggoner & Chen, 2014; Kamble et al., 2015; Caragiannis et al., 2015; Kong et al., 2016; Shnayder et al., 2016).

We are particularly motivated by the peer grading problem, which we will use as a running example. Students benefit from open-ended assignments such as essays or proofs. However, such assignments are used relatively sparingly, particularly in large classes, because they require considerable time and effort to grade properly. An efficient and scalable alternative is having students grade each other (and, in the process, learn from each other's work). Many peer grading systems have been proposed and evaluated in both the artificial intelligence and education literatures (Hamer et al., 2005; Cho & Schunn, 2007; Paré & Joordens, 2008; Shah et al., 2013; de Alfaro & Shavlovsky, 2014; Kulkarni et al., 2014; Raman & Joachims, 2014; Wright et al., 2015; Caragiannis et al., 2015; de Alfaro et al., 2015), albeit with a focus on evaluating the accuracy of grades collected under the assumption of full cooperation by students.

However, no experienced teacher would expect all students to behave nonstrategically when asked to invest effort in a time-consuming task. An effective peer grading system must therefore provide motivation for students to formulate evaluations carefully and to report them honestly. Many approaches have been developed to provide such motivation. One notable category is peer-prediction methods (Prelec, 2004; Miller et al., 2005; Jurca & Faltings, 2009; Faltings et al., 2012; Witkowski & Parkes, 2012; Witkowski et al., 2013; Dasgupta & Ghosh, 2013; Witkowski & Parkes, 2013; Radanovic & Faltings, 2013, 2014; Radanovic et al., 2016; Riley, 2014; Zhang & Chen, 2014; Waggoner & Chen, 2014; Kamble et al., 2015; Kong et al., 2016; Shnayder et al., 2016). In order to motivate each agent to reveal his private, informative signal, peer-prediction methods offer a reward based on how each agent's reports compare with those of his peers. Such rewards are designed to induce truth telling in equilibrium—that is, they create a situation in which each agent has an interest in investing effort and revealing his private and informative signal truthfully, as long as he believes that all other agents will do the same.

Even if they do offer a truthful equilibrium, peer-prediction methods also always induce other uninformative equilibria, the existence of which is inevitable (Jurca & Faltings, 2009; Waggoner & Chen, 2014). Intuitively, if no other agent follows a strategy that depends on her private information, there is no reason for a given agent to deviate in a way that does so either: agents can only be rewarded for coordination, not for accuracy. When private information is costly to obtain, uninformative equilibria are typically *less* demanding for agents to play. This raises significant doubt about whether peer-prediction methods can motivate truthful reporting in practice. Experimental evaluations of peer-prediction methods have mixed results. Some studies showed that agents reported truthfully (Shaw et al., 2011; John et al., 2012; Faltings et al., 2014; Radanovic et al., 2016); another study found that agents colluded on uninformative equilibria (Gao et al., 2014).

Recent progress on peer-prediction mechanisms has focused on making the truthful equilibrium Pareto dominant, i.e., (weakly) more rewarding to every agent than any other equilibrium (Dasgupta & Ghosh, 2013; Witkowski & Parkes, 2013; Kamble et al., 2015; Radanovic & Faltings, 2015; Shnayder et al., 2016). This can be achieved by rewarding agents based on the distributions of their reports for multiple objects. However, we show in this paper that such arguments rely critically on the assumption that every agent has access to only one private signal per object. This is often untrue in practice; e.g., in peer grading, by taking a quick glance at an essay a student can observe characteristics such as length, formatting and the prevalence of grammatical errors. These characteristics require hardly any effort to observe, can be arbitrarily uninformative about true quality, and are of no interest to the mechanism. Yet their existence provides a means for the agents to coordinate. We build on this intuition to prove that no mechanism can guarantee that an equilibrium in which all agents truthfully report their informative signals is always Pareto dominant in every setting.

Motivated by these negative results, we move on to consider a setting in which the operator of the mechanism has access to trusted evaluators (e.g., teaching assistants) who can reliably provide noisy but informative signals of the object's true quality. This allows for a hybrid mechanism that blends peer-prediction with comparison to trusted reports. With a fixed probability, the mechanism obtains a trusted report and rewards the agent based on the agreement between the agent's report and the trusted report (Jurca & Faltings, 2005). Otherwise, the mechanism rewards the agent using a peer-prediction mechanism. Such hybrid mechanisms can yield stronger incentive guarantees than other peer-prediction mechanisms, such as achieving truthful reporting of informative signals in Pareto-dominant equilibrium (see, e.g., Jurca & Faltings, 2005; Dasgupta & Ghosh, 2013). Intuitively, if an agent seeks to be consistently close to a trusted report, then his best strategy is to reveal his informative signal truthfully.

In fact, the availability of trusted reports is so powerful that it gives us the option of dispensing with peer-prediction altogether. Specifically, we can reward students based on agreement with the trusted report when the latter is available, but simply give students a constant reward otherwise, in an approach we dub the *peer-insensitive mechanism*. Indeed, in Wright et al. (2015) we introduced such a peer grading system and showed that it worked effectively in practice, based on a study across three years of a large class. This mechanism has even stronger incentive properties than the hybrid mechanism—because it induces a single-agent game, it can give rise to dominant-strategy truthfulness.

Our paper's main focus is on comparing these two approaches in terms of the number of trusted reports that they require. One might expect that the hybrid approach would have the edge, both because it relies on a weaker solution concept and because it leverages a second source of information reported by other agents. Surprisingly, we prove that this intuition is backwards. We identify a simple sufficient condition, which, if satisfied, guarantees that the peer-insensitive mechanism offers the dominant strategy of truthful reporting of informative signals while querying trusted reports with a lower probability than is required for a peer-prediction mechanism to motivate truthful reporting in Pareto-dominant equilibrium. We then show that all applicable peer-prediction mechanisms of which we are aware satisfy this sufficient condition.

2. Peer-Prediction Mechanisms and Other Related Work

We formally define the game theoretic setting in which we will study the elicitation problem. A mechanism designer wishes to elicit information about the quality of a set O of objects. Each object j has a latent quality $q_j \in Q$, where Q is a finite set. There are n rational and risk-neutral agents. Each agent i evaluates a set of objects J_i, which is a subset of O.

Agents have access to private information about the qualities of the objects of interest, and our goal is to motivate the agents to reveal their private information. To do this, we build upon several peer prediction mechanisms proposed in the literature.

In the peer prediction literature, it is standard to assume that each agent receives a single, private signal, which represents the *only* information that agent has about the object of interest. We argue that, in reality, every agent can obtain multiple pieces of information about the object's quality by investing different amounts of efforts. To capture this, we consider a simplified scenario by assuming that, for each object j, agent i has access to two pieces of private information: a *high-quality signal* $s^h \in Q$ and a *low-quality signal* s^l.

The high-quality signal refers to a useful piece of information that the mechanism designer wishes to elicit. It is drawn from a distribution conditional on the object's quality q_j and the identity of the agents evaluating the object. The joint distributions of the high-quality signals are common knowledge among the agents. The high-quality signals can be arbitrarily correlated with one another. In particular, we do not assume that the high-quality signals are independent conditional on the object's quality and the identity of the agents. Agent i can form a probabilistic belief about the high-quality signal of another agent i' by performing a Bayesian update based on his own high-quality signal.

The low-quality signal represents information about superficial qualities of the object — it is correlated with the quality of the object, but not sufficiently so. The mechanism designer prefers to get the high-quality signals rather than the low-quality signals because the high-quality signals are more correlated with the quality of the object than the low-quality signals. However, the low-quality signals are easier to obtain than the high-quality signals because the low-quality signals of different agents are more correlated with each other than the high-quality signals of these agents. As a result, the low-quality signals provide an easier way for agents to coordinate their reports compared to investing costly effort to acquire the high-quality signal.

The low-quality signal is a collection of properties of the object that an agent can observe with negligible effort. For example, by glancing at an essay and skimming several sentences, an agent can observe several superficial attributes of the essay, such as the length of the essay, whether the author provided references or not, the number of spelling and grammatical mistakes, the sentence structure, the vocabulary, and the complexity of the language being used[1]. Similarly, one could base a review on the decor without eating in a restaurant; evaluate a movie by watching its trailer; etc.

In practice, it is often costly to perform careful evaluations by obtaining the high-quality signals. We capture this by assuming that obtaining the high-quality signal requires a constant effort $c^E > 0$ whereas obtaining the low-quality signal requires no effort. Our results can be straightforwardly extended to cases where the difference between the cost of obtaining the two signals is c^E.

We consider mechanisms that may ask each agent to report up to two pieces of information. For each object j evaluated, agent i makes a signal report $r_{ij} \in Q$ and a belief report $b_{ij} \in \Delta(Q)$ to the mechanism, where $\Delta(Q)$ is the set of all distributions over Q. The mechanism gives agent i a reward $z_{ij}(\mathbf{r}, \mathbf{b})$ based on the vector of all the signal reports $\mathbf{r}$ and belief reports $\mathbf{b}$ for object j.

Agents may strategize over both whether to incur the cost of effort to observe the high-quality signal and which signal and belief reports to make. The goal of the mechanism designer is to incentivize each agent i to incur the cost of effort to obtain the high-quality signal s^h_{ij}, to report the high-quality signal as his signal report, and to report his posterior belief conditional on the high-quality signal as his belief report. A mechanism has a *truthful equilibrium* when it is a Bayesian Nash equilibrium equilibrium for every agent to use this strategy.

We say that a peer-prediction mechanism is *universal* if it can be applied without prior knowledge of the

[1] In a large class, it is unlikely that all the students could agree on which superficial attribute of an essay they should coordinate on. However, by combining multiple superficial attributes together, the students could arrive at similar estimates of the essay, which is the low-quality signal.

distribution from which signals are elicited, and for any number of agents greater than or equal to 3:

Definition 1 (Universal peer-prediction mechanism). *A peer-prediction mechanism is* universal *if it can be operated without knowledge of the joint distribution of the high-quality signals s_{ij}^h (i.e., it is "detail free" (Wilson, 1987)) and guarantees the existence of the truthful equilibrium for any number of agents $n \geq 3$ and any number of tasks.*

We focus on universal peer-prediction mechanisms for two reasons. First, in practice, it is not always possible for a mechanism designer to have detailed knowledge of the joint signal distribution, so this allows us to focus on mechanisms that are more likely to be used in practice. Second, it is relatively unrestrictive, as many peer-prediction mechanisms in the literature satisfy universality.

Below, we give a high-level description of universal and non-universal peer-prediction mechanisms as well as other related work. We first introduce universal peer-prediction mechanisms, which can be divided into three categories: output agreement mechanisms, multi-object mechanisms, and belief based mechanisms.

Output-Agreement Mechanisms Output agreement mechanisms were first introduced by Von Ahn & Dabbish (2008) and later studied by Faltings et al. (2012); Witkowski et al. (2013); Waggoner & Chen (2014).

Output agreement mechanisms only collect signal reports from agents and reward an agent i for evaluating object j based on agents' signal reports for the object (Faltings et al., 2012; Witkowski et al., 2013; Waggoner & Chen, 2014).

The standard output agreement mechanism studied by Waggoner & Chen (2014) and Witkowski et al. (2013) gives an agent i a constant reward exactly when agent i's signal report matches the signal report of another random agent i' evaluating the same object. The Faltings et al. (2012) mechanism also rewards the agents for agreement, but the amount of the reward is scaled by the empirical frequency of the signal report agreed upon. The more frequently the report appears, the smaller the reward.

Multi-Object Mechanisms Multi-object mechanisms reward each agent based on his reports for multiple objects (Dasgupta & Ghosh, 2013; Radanovic & Faltings, 2015; Kamble et al., 2015; Shnayder et al., 2016).

The multi-signal Dasgupta-Ghosh mechanism (Dasgupta & Ghosh, 2013; Shnayder et al., 2016) and the Kamble et al. (2015) mechanism extend the output agreement mechanisms by adding additional scaling terms to the reward. The Shnayder et al. (2016) mechanism adds an additive scaling term, whereas the Kamble et al. (2015) mechanism adds a multiplicative scaling term. These scaling terms are intended to exploit correlations between multiple tasks to make the truthful equilibrium dominate a particular kind of uninformative equilibria, by reducing the reward to agents who agree to a report that is "unsurprising" given their reports on other objects.

The Radanovic & Faltings (2015) mechanism rewards the agents for report agreement using a reward function inspired by the quadratic scoring rule. A quadratic scoring rule is a proper scoring rule, which is designed to incentivize an agent to report his belief about the likelihoods of the outcomes of an event truthfully.

Belief Based Mechanisms Belief based mechanisms collect both signal and belief reports from agents and reward each agent based on all agents' signal and belief reports for each object (Witkowski & Parkes, 2012, 2013; Radanovic & Faltings, 2013, 2014; Riley, 2014). These mechanisms make use of proper scoring rules, which are designed to incentivize an agent to report his belief truthfully.

The robust Bayesian Truth Serum (BTS) (Witkowski & Parkes, 2012, 2013) rewards agent i for evaluating object j by how well his belief report b_{ij} and shadowed belief report b^s_{ij} predict the signal reports of another random agent i''. Agent i's shadowed belief report b^s_{ij} is the result of modifying another agent i''s belief report based on agent i's signal report.

The multi-valued robust BTS (Radanovic & Faltings, 2013) rewards agent i if his signal report matches that of another random agent i' and his belief report accurately predicts agent i''s signal report.

The divergence-based BTS (Radanovic & Faltings, 2014) rewards agent i for evaluating object j if his belief report accurately predicts another random agent's signal report. Moreover, it penalizes agent i if his signal report matches the signal report of another agent i' but his belief report is sufficiently different from the belief report of agent i'.

The Riley (2014) mechanism rewards agent i for evaluating object j by how well his belief report predicts other agents' signal reports. Moreover, agent i's reward is bounded above by the score for the average belief report of other agents who made the same signal report.

Non-Universal Peer-Prediction Mechanisms We do not consider several peer-prediction mechanisms because they are not universal according to Definition 1. The Miller et al. (2005); Zhang & Chen (2014) and Kong et al. (2016) mechanisms require the mechanism operator to derive the agents' posterior beliefs based on their signal reports (hence requiring knowledge of the distribution from which signals are drawn); they all then reward the agents based on how well the derived posterior belief predicts other agents' signal reports using proper scoring rules. The Jurca & Faltings (2009) mechanism requires the joint distribution of the signals to construct rewards that either penalize or eliminate symmetric, uninformative equilibria. The Correlated Agreement mechanism (Shnayder et al., 2016) requires the joint distribution of the signals to determine the rewards. The Correlated Agreement Detail-Free mechanism (Shnayder et al., 2016) requires a large number of tasks to guarantee that the truthful equilibrium exists and yields higher expected payment than any other strategy profile. The Bayesian Truth Serum (BTS) mechanism (Prelec, 2004) requires an infinite number of agents to guarantee the existence of the truthful equilibrium. We note that Prelec (2004) pioneered the idea of eliciting both signal and belief reports from agents. Much subsequent work leveraged this key idea to sustain the truthful equilibrium while not requiring knowledge of the prior distributions of the signals to operate the mechanism (Witkowski & Parkes, 2012, 2013; Radanovic & Faltings, 2013, 2014;

Riley, 2014).

In a follow-up work, Liu & Chen (2018) proposed a peer prediction mechanism called Dominant Truth Serum (DTS). We characterize DTS as a non-universal mechanism for the following reasons. In a setting with multiple agents and multiple tasks, DTS guarantees the existence of a truthful equilibrium if the mechanism knows the average error rates in the agents' truthful reports. The mechanism can obtain such knowledge either by knowing part of the joint distribution of the agents' signals or by estimating the error rates from data. In the latter case, the estimates are accurate only if the number of agents and the number of tasks per agent are large enough.

Hierarchical Mechanism (de Alfaro et al., 2015) Independent to our work, de Alfaro et al. (2015) also proposed the idea of using peer prediction mechanisms in conjunction with limited access to trusted reports. In their hierarchical mechanism, students are placed into a tree structure. Students in the top layer of the tree are incentivized through trusted reports whereas students in the layers below are incentivized via a peer prediction mechanism. By an inductive argument, the truthful equilibrium exists and is unique, so long as the top-layer students are sufficiently incentivized. This mechanism is detail free with respect to the distribution of signals, and is thus universal. However, the existence of the truthful equilibrium requires every student to know which layer of the tree structure they occupy; that is, different students are treated differently ex-ante. In this work, the anonymity assumption is violated, and this violation turns out to have major implications for the properties of the mechanism.

In future work we intend to further explore relaxations of the single-signal assumption and anonymity, and connections between them.

Other Related Work Other work in the broader artificial intelligence literature also considers eliciting truthful reports from self-interested agents. However, this work assumes that the information will eventually be costlessly revealed to the center, unlike our setting where every observation of ground truth is costly.

Papakonstantinou et al. (2008, 2010, 2011) study the problem of eliciting costly probabilistic estimates, with a specified minimum precision, from one or more agents. However, they assume that the center has no knowledge of agents' costs of acquiring their estimates, whereas in our work we assume that agents' costs are known by the center. They propose a two-stage mechanism, which elicits agents' true costs and incentivizes agents to truthfully report their estimates.

Zohar & Rosenschein (2006, 2008) study the problem of eliciting private information from agents when agents have different beliefs about the underlying probability of events, either due to differing priors, or due to the agents' being able to observe a costless signal before deciding whether to observe their costly signal. Unlike our setting, the center eventually receives a noisy signal of ground truth without cost. They propose efficient algorithms to construct mechanisms that are robust to small differences between the agents' beliefs and the center's. When the agents are not willing to share certain information they possess, however, they show that designing an optimal mechanism becomes computationally hard. They also propose mechanisms that elicit agents' confidence about their information in addition to the information itself.

3. Impossibility of Pareto-Dominant, Truthful Elicitation

In this section, we show that when agents have access to multiple signals about an object, Pareto-dominant truthful elicitation is impossible for any universal mechanism that computes agent rewards solely based on agents' reports without any access to ground truth. Intuitively, without knowledge of the distributions from which the signals are drawn, the mechanism cannot distinguish the signal that it hopes to elicit from other, irrelevant signals.

We consider universal mechanisms that compute agent rewards solely based on agents' reports. We define a *multi-signal environment* formally below.

Definition 2. *A* multi-signal environment *is a setting in which agents have access to at least two signals, the high quality signal and a low-quality signal.*

Recall that the obtaining the high-quality signal requires a constant amount of effort $c^E > 0$, whereas obtaining the low-quality signal requires no effort. Let π_i^s denote agent i's ex-ante expected reward at an equilibrium of the mechanism at which every agent reports the s signal truthfully. Let the *truthful equilibrium* refer to the equilibrium in which each agent reports their high-quality signal truthfully.

We care about developing mechanisms for which the truthful equilibrium is Pareto dominant. We define Pareto dominance below.

Definition 3. *The e_1 equilibrium* Pareto dominates *the e_2 equilibrium if and only if every agent's expected utility at e_1 is greater than or equal to his expected utility at e_2 and there is at least one agent whose expected utility at e_1 is greater than his expected utility at e_2.*

The e_1 equilibrium is Pareto dominant *if it Pareto dominates every other equilibrium of the mechanism.*

Theorem 1. *For any universal mechanism, if obtaining the high-quality signals requires an additional cost of effort of $c > 0$ compared to obtaining any low-quality signal, then there exists a multi-signal environment in which the truthful equilibrium is not Pareto dominant.*

Proof. Consider a universal mechanism. For any signal s, let the s equilibrium be the equilibrium in which all agents report the signal s truthfully.

Based on Definition 2, consider a multi-signal environment in which agents have access to two signals s and s'. Assume that there is an equilibrium of the mechanism at which all agents truthfully report the s signal and that there is another equilibrium at which every agent reports the s' signal truthfully. If one of these equilibria does not exist, then the theorem is trivially true for the case where the corresponding signal is the high-quality signal. Consider three cases.

Case 1: s is the high-quality signal and the s equilibrium Pareto dominates the s' equilibrium. This means that $\pi_i^s - c > \pi_i^{s'}$ for some agent i and $\pi_j^s - c \geq \pi_j^{s'}$ for all agents $j \neq i$. In this case, consider another multi-signal environment with signals s and s', with identical joint distribution as s and s' in the original environment, but in which s' is the high-quality signal and s is the low-quality signal. The theorem holds in this environment because the truthful (s') equilibrium is Pareto dominated by the s equilibrium.

Case 2: s' is the high-quality signal and the s' equilibrium Pareto dominates the s equilibrium. Using a similar argument as in case 1, we can construct a multi-signal environment in which s is the high-quality signal and s' is the low-quality signal. This theorem holds in this environment.

Case 3: Neither of the first two cases is true. Let s be the high-quality signal. Since case 1 is not satisfied, the truthful (s) equilibrium does not Pareto dominate the s' equilibrium. Therefore, the theorem holds in this environment. □

Any universal mechanism does not have access to the joint distributions of the signal. Therefore, with multiple signals, there is no way for a universal mechanism to ensure that the truthful equilibrium yields the highest utility for the agents. The truthful equilibrium is Pareto dominant only if the high-quality signal happens to be drawn from a distribution yielding sufficiently higher reward than every other signal to compensate for the cost of effort.

One way for the mechanism designer to ensure that agents are reporting the high-quality signal is to stochastically compare agents' reports to ground truth. In the next section, we introduce a class of mechanisms that takes this approach.

4. Combining Elicitation with Limited Access to Ground Truth

Elicitation mechanisms are designed for situations where it is infeasible for the mechanism designer to evaluate each object herself. However, in practice, it is virtually always possible, albeit costly, to obtain *trusted reports*, i.e., unbiased evaluations of a subset of the objects. In the peer grading setting, the instructor and teaching assistants can always mark some of the assignments. Similarly, review sites could in principle hire an expert to evaluate restaurants or hotels that its users have reviewed; and so on.

In this section, we define a class of mechanisms that take advantage of this limited access to ground truth to circumvent the result from Section 3. The mechanism performs a spot check on each object with some probability. When a spot check is performed, the mechanism obtains a trusted report s_j^t, which is an unbiased estimator of the object's quality, and rewards each agent by comparing the agent's signal and belief reports with the trusted report [2]. We define such a spot-checking mechanism as follows.

Definition 4 (spot-checking mechanism)**.** *A* spot-checking mechanism *is a tuple $M = (p, y_{ij}, z_{ij})$, where p is the* spot-check probability*; $y_{ij}(r_{ij}, s_j^t)$ is a* checked payment rule*; and $z_{ij}(\boldsymbol{r}, \boldsymbol{b})$ is an* unchecked payment rule. *With probability p, the mechanism rewards the agent based on his signal report r_{ij} and the trusted report s_j^t according to the spot-check mechanism y_{ij}. With probability $1-p$, the mechanism rewards the agent according to the unchecked payment rule z_{ij}.*

Formally, agent i's reward for evaluating object j is given by

$$\pi_{ij} = \begin{cases} y_{ij}(r_{ij}, s_j^t) & \textit{if object } j \textit{ is spot checked,} \\ z_{ij}(\boldsymbol{r}, \boldsymbol{b}) & \textit{otherwise.} \end{cases} \tag{1}$$

In this work we compare two approaches to using limited access to ground truth for elicitation. The first approach is to augment an existing peer-prediction mechanism with spot-checking:

Definition 5 (spot-checking peer-prediction mechanism)**.** *A spot-checking peer prediction mechanism is a spot-checking mechanism (p, y_{ij}, z_{ij}) in which the unchecked payment rule z_{ij} is a peer prediction mechanism.*

The second approach is to rely exclusively on ground truth access to incentivize truthful reporting:

Definition 6 (peer-insensitive mechanism)**.** *A* peer-insensitive mechanism *is a spot-checking mechanism (p, y_{ij}, z_{ij}) in which the unchecked payment rule is a constant function. That is, $z_{ij}(\boldsymbol{r}, \boldsymbol{b}) = W$ for some constant $W > 0$.*

We assume that the mechanism designer has no value for the reward given to the agents. Instead, we seek only to minimize the spot-check probability required to make the truthful equilibrium either unique or Pareto dominant, since access to trusted reports is assumed to be costly.[3] This models situations where agents are rewarded by grades (as in peer grading), virtual points or badges (as in online reviews), or other artificial currencies.

We fix the checked payment rule as defined in Equation (2), using a form inspired by the multi-signal Dasgupta-Ghosh mechanism (Dasgupta & Ghosh, 2013; Shnayder et al., 2016). Let J^t be the set of objects that was spot-checked. Let object j be evaluated by agent i and be spot checked. Let $j' \in J_i$ be an object evaluated by agent i, chosen uniformly at random among all the objects evaluated by agent i. Let $j'' \in J^t \backslash J_i$ be an object that was spot checked, also chosen uniformly at random among all the objects spot

[2] For each object being evaluated, the mechanism needs to obtain at most one trusted report. In the peer grading setting, students evaluate multiple submissions. Therefore, the mechanism may need to obtain multiple trusted reports overall.

[3] If access to trusted reports were not costly, then querying strategic agents rather than trusted reports on all the objects would be pointless.

checked.[4] Then agent i's reward for object j is

$$y_{ij}(\mathbf{r}, \mathbf{s^t}) = \mathbb{1}_{r_{ij}=s^t_j} - \mathbb{1}_{r_{ij'}=s^t_{j''}}. \tag{2}$$

Agents may strategically choose whether or not to incur the cost of observing the high-quality signal, and having chosen which signal to observe, may report any function of either signal. Formally, let G^h_i be the set of all full-effort pure strategies, where an agent observes the high-quality signal—incurring observation cost c^E—and then reports a function $g(s^h_{ij})$ of the observed value. Let G^l_i be the set of all no-effort pure strategies, where an agent observes the low-quality signal—incurring no observation cost—and then reports a function $g(s^l_{ij})$ of the observed value. The set of pure strategies available to an agent is thus $G^h_i \cup G^l_i$. We assume that agents apply the same strategy to every object that they evaluate; however, we allow agents to play a mixed strategy by choosing the mapping stochastically.

We define the g^l strategy to be an agent's best no-effort strategy when a spot check is performed. What is special about this strategy is that, if an agent chooses to invest no effort, then this is his best strategy for any spot-check probability $p \in [0, 1]$. Thus, the g^l equilibrium is the best equilibrium for all agents conditional on investing no effort.

Definition 7. *Let $g^l = \arg\max_{g \in G^l} \mathbb{E}[y(g(s^l), s^t)]$ be an agent's best no-effort strategy when a spot check is performed. Let the g^l equilibrium be the equilibrium where every agent uses the g^l strategy if such an equilibrium exists.*

We assume that the high-quality and the low-quality signals are both categorical with respect to the trusted report. A signal is categorical if, when an agent observes a realization of the signal, all other realizations of the trusted report become less likely than their prior probabilities. Formally,

Definition 8 (categorical signals)**.** *The low-quality signal s^l is* categorical *if and only if* $\Pr(s^t = s' | s^l = s) < \Pr(s^t = s')$*, for all $s, s' \in Q$ and $s \neq s'$. The high-quality signal s^h is categorical if and only if* $\Pr(s^t = s' | s^h = s) < \Pr(s^t = s')$*, for all $s, s' \in Q$ and $s \neq s'$.*

The categorical assumption implies that each type of signal is positively correlated with the trusted report. This assumption is important to ensure that comparing with a trusted report is sufficient to incentivize an agent to obtain the high-quality or the low-quality signal given that the agents invests full or no effort. With categorical signals, we can show that, if an agent invests full effort, then he maximizes his spot-check reward by obtaining the high-quality signal and reporting it truthfully. Similarly, if an agent invests no effort, he maximizes his spot-check reward by obtaining the low-quality signal and reporting it truthfully.

[4] Note that in Dasgupta & Ghosh (2013), it is important for strategic reasons that object j' has not been evaluated by the opposing agent; this is not important in our setting, since the trusted reports are assumed to be nonstrategic.

We assume that each type of signal satisfies the assumptions to ensure the existence of the truthful equilibrium when the spot-check probability is 0 for every peer prediction mechanism we consider in Corollaries 1 and 2. Note that the categorical assumption is sufficient to ensure the existence of the truthful equilibrium for output-agreement peer prediction mechanisms.

In practice, it is reasonable to assume that coordinating on the low-quality signal yields more payoff than coordinating on the high-quality signal even when no spot check is performed and the high-quality signal requires no effort to obtain. We capture this by making additional assumptions about the low-quality signal. First, we assume that the low-quality signal is perfectly correlated across agents. Since the high-quality signal is noisy, this assumption implies that the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$ for every universal peer prediction mechanism that we consider.

Assumption 1. *The low-quality signal is perfectly correlated across all the agents.*

We also assume that the low-quality signal s^l is drawn from a uniform distribution over Q; this is essentially without loss of generality, since in any setting where the agents see a description of the object as well as their evaluation, a uniform distribution can be obtained by, e.g., hashing the description. More realistically, objects may have names or lengths that are approximately uniformly distributed. This assumption ensures that the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$ for the multi-signal Dasgupta-Ghosh mechanism (Dasgupta & Ghosh, 2013; Shnayder et al., 2016) and the Kamble et al. (2015) mechanism.[5]

Assumption 2. *For any agent and any object, the low-quality signal is drawn from a uniform distribution over the set of qualities Q.*

We assume that the correlation between the high-quality signal and the object's quality is sufficiently high to compensate for the cost of effort, even though agents have the option of getting the low-quality signals at no cost. In other words, when the object is spot checked, paying the cost of observing the high-quality signal is worthwhile. Formally,

Assumption 3. $\mathbb{E}\left[y_{ij}(s^h, s^t) - c^E\right] > \mathbb{E}\left[y_{ij}(g^l(s^l), s^t)\right]$ *for any agent i and object j.*

This assumption implies that, when the object is spot checked, an agent prefers to pay the cost to observe and report the high-quality signal rather than playing the best strategy conditional on observing the low-quality signal for free. As an extreme example, if the low-quality signal were perfectly correlated with the quality, then no amount of spot-checking would induce an agent to observe the high-quality signal (nor, indeed, would a mechanism designer want them to).

[5] We could also derive a weaker assumption for each individual peer prediction mechanism that implies the above property.

One consequence of Assumption 3 is that scaling the rewards would not be sufficient to incentivize the agents to obtain the high-quality signals. With scaling, the reward for obtaining the high-quality and the low-quality signals would both increase. Agents would still prefer to obtain the low-quality signals if they are not spot-checked.

5. When Does Peer Prediction Help?

We compare the peer-insensitive mechanism with all universal spot-checking peer-prediction mechanisms. Theorem 2 states that, if a simple sufficient condition is satisfied, then compared to all universal spot-checking peer-prediction mechanisms, the peer-insensitive mechanism can achieve stronger incentive properties (dominant-strategy truthfulness versus Pareto dominance of truthful equilibrium) while requiring a smaller spot-check probability.

In Lemma 1, we derive an expression for the minimum spot-check probability p_{ds} at which the truthful strategy is a dominant strategy for the peer-insensitive mechanism. When the spot-check probability is p_{ds}, any agent is indifferent between playing the g^l strategy and investing effort and reporting truthfully. Recall that the g^l strategy is an agent's best strategy conditional on investing no effort when the object is spot checked.

Lemma 1. *The minimum spot-check probability p_{ds} at which the truthful strategy is dominant for the peer-insensitive mechanism satisfies the following equation.*

$$p_{\text{ds}}\,\mathbb{E}[y(s^h, s^t)] - c^E = p_{\text{ds}}\,\mathbb{E}[y(g^l(s^l), s^t)]. \tag{3}$$

Proof. Please see Appendix A. □

Next, we consider any spot-checking peer-prediction mechanism. Our goal is to derive a lower bound for p_{Pareto}, the minimum spot-check probability at which the truthful equilibrium is Pareto dominant.

For the truthful equilibrium to be Pareto dominant, it is necessary that the truthful equilibrium Pareto dominates the g^l equilibrium. There are two ways to make the truthful equilibrium Pareto dominate the g^l equilibrium. If we increase the spot-check probability until the g^l equilibrium is eliminated, then the truthful equilibrium trivially Pareto dominates the g^l equilibrium. Let p_{el} denote the minimum spot-check probability at which the g^l equilibrium is eliminated. Otherwise, we can increase the spot-check probability to a value at which the truthful equilibrium Pareto dominates the g^l equilibrium assuming that the g^l equilibrium exists at this spot-check probability. Let p_{ex} denote the minimum spot-check probability at which an agent receives higher expected utility at the truthful equilibrium than at the g^l equilibrium, assuming that the g^l equilibrium exists when the spot-check probability is p_{ex}. The minimum of p_{el} and p_{ex} is the minimum spot-check probability at which the truthful equilibrium Pareto dominates the g^l equilibrium, and it is also a lower bound for p_{Pareto}.

In Lemma 2, we characterize the minimum spot-check probability p_{el} at which the g^l equilibrium is eliminated, and we show that p_{el} is greater than the minimum spot-check probability p_{ds} to motivate a single agent to report truthfully, under certain assumptions. To eliminate the g^l equilibrium, we need to increase the spot-check probability enough such that an agent prefers to play his best strategy with full effort rather than playing the g^l strategy while all other agents follow the g^l equilibrium. Persuading an agent to deviate from the g^l equilibrium is difficult for two reasons. First, an agent incurs a cost by deviating from the g^l equilibrium when all other agents follow it. Second, the agent's best strategy with full effort gives him no greater spot-check reward than the truthful strategy. The combined effect means that we need a higher spot-check probability to persuade an agent to deviate from the g^l equilibrium than to motivate a single agent to report truthfully.

The sufficient conditions characterized in Lemma 2, Lemma 3 and Theorem 2 hold whenever $c^E = 0$. Moreover, we will show that all universal peer-prediction mechanisms in the literature satisfy these sufficient conditions for all $c^E \geq 0$.

Lemma 2 states that p_{el} is greater than or equal to p_{ds} under certain assumptions. Intuitively, this means that, we need a higher spot-check probability to make the eliminate the g^l equilibrium than to motivate a single agent to report truthfully. When the g^l equilibrium is eliminated, the truthful equilibrium trivially Pareto dominates the g^l equilibrium.

Lemma 2. *For any spot-checking peer-prediction mechanism, if the g^l equilibrium exists when $c^E = 0$ and $p = 0$, then $p_{\mathrm{el}} \geq p_{\mathrm{ds}}$ for all $c^E \geq 0$.*

Proof. Please see Appendix B. □

Lemma 3 states that p_{ex} is greater than or equal to p_{ds} under certain assumptions. The intuition is that, when no spot check is performed, the g^l equilibrium Pareto dominates the truthful equilibrium. Thus, assuming that the g^l equilibrium exists, we need a higher spot-check probability to make the truthful equilibrium Pareto dominate the g^l equilibrium than to motivate a single agent to report truthfully.

Lemma 3. *For any spot-checking peer-prediction mechanism, if the g^l equilibrium exists and Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$, then $p_{\mathrm{ex}} \geq p_{\mathrm{ds}}$ for all $c^E \geq 0$.*

Proof. Please see Appendix C. □

If the conditions in Lemmas 2 and 3 hold, then the minimum of p_{el} and p_{ex} is greater than or equal to p_{ds}. Since the minimum of p_{el} and p_{ex} is a lower bound of p_{Pareto}, it must be that $p_{\mathrm{Pareto}} \geq p_{\mathrm{ds}}$. In Theorem 2, we prove that the conditions in Lemmas 2 and 3 are sufficient conditions for $p_{\mathrm{Pareto}} \geq p_{\mathrm{ds}}$ — the minimum spot-check probability to make the truthful equilibrium Pareto dominant for a spot-checking peer-prediction mechanism is higher than the spot-check probability to make the truthful strategy dominant for the peer-insensitive mechanism.

Theorem 2 (Sufficient condition for Pareto comparison)**.** *For any spot-checking peer-prediction mechanism, if the g^l equilibrium exists and Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$, then $p_{\text{Pareto}} \geq p_{\text{ds}}$ for all $c^E \geq 0$.*

Proof. Please see Appendix D. □

We now show that, under very natural conditions, *every* universal peer-prediction mechanism of which we are aware in the literature satisfies the conditions of Theorem 2; hence, in this setting, the peer-insensitive spot-checking mechanism requires less ground truth access than any spot-checking peer-prediction mechanism.

Corollary 1. *For spot-checking peer-prediction mechanisms based on Faltings et al. (2012); Witkowski et al. (2013); Dasgupta & Ghosh (2013); Waggoner & Chen (2014); Kamble et al. (2015); Radanovic & Faltings (2015) and Shnayder et al. (2016), the minimum spot-check probability p_{Pareto} for the Pareto dominance of the truthful equilibrium is greater than or equal to the minimum spot-check probability p_{ds} at which the truthful strategy is a dominant strategy for the peer-insensitive mechanism.*

Proof. Please see Appendix F. □

Corollary 2. *For spot-checking peer-prediction mechanisms based on Witkowski & Parkes (2012, 2013); Radanovic & Faltings (2013, 2014) and Riley (2014), if the peer-prediction mechanism uses a symmetric proper scoring rule, then the minimum spot-check probability p_{Pareto} for the Pareto dominance of the truthful equilibrium is greater than or equal to the minimum spot-check probability p_{ds} at which the truthful strategy is a dominant strategy for the peer-insensitive mechanism.*

Proof. Please see Appendix G. □

To prove Corollaries 1 and 2, it suffices to show that every mechanism in the corollaries satisfies the sufficient conditions in Theorem 2. To do this, we first need to determine what the g^l strategy is. In other words, what is an agent's best strategy conditional on being spot checked and investing no effort? In Appendix F, we prove in Lemma 4 that the g^l strategy is to report the low-quality signal s^l truthfully. Given Lemma 4, in the proofs of the two corollaries, we show that, for every mechanism considered, the s^l equilibrium exists and Pareto dominates the truthful equilibrium when the cost of effort and the spot-check probability are both zero.

In proving Corollary 1, we made use of two key insights. First, the mechanisms reward agents for agreeing on their reports. Thus, coordinating on reporting the low-quality signal s^l is an equilibrium. Second, the low-quality signal s^l is noiseless whereas the high-quality signal s^h is noisy. As a result, playing the s^l equilibrium yields a higher degree of agreement and higher expected utilities for the agents than playing the s^h equilibrium.

In Corollary 2, we consider spot-checking peer-prediction mechanisms which ask agents to provide belief reports in additional to signal reports. For such mechanisms, in addition to the ideas used to prove Corollary 1, we need an additional assumption that the proper scoring rule used by the mechanism is symmetric. For a symmetric scoring rule, the expected score for reporting a signal s and reporting a belief b_s which predicts that the signal s is observed with probability 1 is the same for every signal $s \in Q$. As a result, an agent can maximize his expected score by reporting s and b_s for any signal $s \in Q$. This insight is crucial for showing that the s^l equilibrium Pareto dominates the truthful equilibrium for every belief-based mechanism in Corollary 2.

6. Conclusions and Future Work

We consider the problem of using limited access to noisy but unbiased ground truth to incentivize agents to invest costly effort in evaluating and truthfully reporting the quality of some object of interest. Absent such spot-checking, peer-prediction mechanisms already guarantee the existence of a truthful equilibrium that induces both effort and honesty from the agents. However, this truthful equilibrium may be less attractive to the agents than other, uninformative equilibria.

Some mechanisms in the literature have been carefully designed to ensure that the truthful equilibrium is the most attractive equilibrium to the agents (i.e., Pareto dominates all other equilibria). However, these mechanisms rely crucially on the unrealistic assumption that agents' only means of correlating are via the signals that the mechanism aims to elicit. We show that under the more realistic assumption that agents have access to more than one signal, no universal peer-prediction mechanism has a Pareto-dominant truthful equilibrium in all settings.

In contrast, we present a simpler peer-insensitive mechanism that provides incentives for effort and honesty only by checking the agents' reports against ground truth. While one might have expected that peer-prediction would require less frequent access to ground truth to achieve stronger incentive properties than the peer-insensitive mechanism, we proved the opposite for all universal spot-checking peer-prediction mechanisms.

This surprising finding is intuitive in retrospect. Peer-prediction mechanisms can only motivate agents to behave in a certain way as a group. An agent has a strong incentive to be truthful if all other agents are truthful; conversely, when all other agents coordinate on investing no effort, the agent again has a strong incentive to coordinate with the group. Peer-prediction mechanisms thus need to provide a strong enough incentive for agents to deviate from the most attractive uninformative equilibrium in the worst case, whereas the peer-insensitive mechanism only needs to motivate effort and honesty in an effectively single-agent setting.

Many exciting future directions remain to be explored. For example, we assumed that the principal does not care about the total amount of the artificial currency rewarded to the agents. One possible direction would consider a setting in which the principal seeks to minimize both spot checks and the agents' rewards. Also, in our analysis, we assumed that the spot-check probability does not depend on the agents' reports. Conditioning the spot-check probability on the agents' reports might allow the mechanism to more efficiently detect and punish uninformative equilibria. We are particularly excited about designing more sophisticated spot check mechanisms where the spot-check probability is a function of the set of reports for a particular submission. In addition, we are interested in exploring the scenario in which some agents are altruistic and always invest the effort to obtain the high-quality signal.

7. Acknowledgements

Xi Alice Gao was supported by a Postdoctoral Fellowship from the Natural Sciences and Engineering Research Council of Canada. Kevin Leyton-Brown was supported by a Natural Sciences and Engineering Research Council of Canada E.W.R. Steacie Fellowship, Collaborative Research and Development grant, and Discovery Grant, and by a Google Faculty Research Award.

References

de Alfaro, L., Polychronopoulos, V., & Shavlovsky, M. (2015). Incentives for truthful peer grading. *UC Santa Cruz Technical Report*, .

de Alfaro, L., & Shavlovsky, M. (2014). Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education* (pp. 415–420).

Caragiannis, I., Krimpas, G. A., & Voudouris, A. A. (2015). Aggregating partial rankings with applications to peer grading in massive online open courses. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems* (pp. 675–683). International Foundation for Autonomous Agents and Multiagent Systems.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, *48*, 409–426.

Dasgupta, A., & Ghosh, A. (2013). Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on the World Wide Web* (pp. 319–330).

Faltings, B., Jurca, R., Pu, P., & Tran, B. D. (2014). Incentives to counter bias in human computation. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Faltings, B., Li, J. J., & Jurca, R. (2012). Eliciting truthful measurements from a community of sensors. In *3rd International Conference on the Internet of Things (IOT)* (pp. 47–54). IEEE.

Gao, X. A., Mao, A., Chen, Y., & Adams, R. P. (2014). Trick or treat: putting peer prediction to the test. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation* (pp. 507–524). ACM.

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

Hamer, J., Ma, K. T., & Kwong, H. H. (2005). A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian Conference on Computing Education* (pp. 67–72). Australian Computer Society, Inc. volume 42.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, (p. 0956797611430953).

Jurca, R., & Faltings, B. (2005). Enforcing truthful strategies in incentive compatible reputation mechanisms. *Internet and Network Economics*, (pp. 268–277).

Jurca, R., & Faltings, B. (2009). Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, *34*, 209.

Kamble, V., Shah, N., Marn, D., Parekh, A., & Ramachandran, K. (2015). Truth serums for massively crowdsourced evaluation tasks. *arXiv preprint*, *arXiv:1507.07045*.

Kong, Y., Ligett, K., & Schoenebeck, G. (2016). Putting peer prediction under the micro(economic)scope and making truth-telling focal. *arXiv preprint*, *arXiv:1603.07319*.

Kulkarni, C. E., Socher, R., Bernstein, M. S., & Klemmer, S. R. (2014). Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the First ACM Conference on Learning @Scale* (pp. 99–108).

Liu, Y., & Chen, Y. (2018). Surrogate scoring rules and a dominant truth serum for information elicitation. *arXiv preprint arXiv:1802.09158*, .

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

Papakonstantinou, A., Rogers, A., Gerding, E. H., & Jennings, N. R. (2008). A truthful two-stage mechanism for eliciting probabilistic estimates with unknown costs. In *ECAI* (pp. 448–452).

Papakonstantinou, A., Rogers, A., Gerding, E. H., & Jennings, N. R. (2010). Mechanism design for eliciting probabilistic estimates from multiple suppliers with unknown costs and limited precision. In *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets* (pp. 102–116). Springer.

Papakonstantinou, A., Rogers, A., Gerding, E. H., & Jennings, N. R. (2011). Mechanism design for the truthful elicitation of costly probabilistic estimates in distributed information systems. *Artificial Intelligence*, *175*, 648–672.

Paré, D. E., & Joordens, S. (2008). Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *Journal of Computer Assisted Learning*, *24*, 526–540.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*, 462–466.

Radanovic, G., & Faltings, B. (2013). A robust Bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013* (pp. 833–839).

Radanovic, G., & Faltings, B. (2014). Incentives for truthful information elicitation of continuous signals. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Radanovic, G., & Faltings, B. (2015). Incentives for subjective evaluations with private beliefs. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Radanovic, G., Faltings, B., & Jurca, R. (2016). Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *7*, 48.

Raman, K., & Joachims, T. (2014). Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1037–1046). ACM.

Riley, B. (2014). Minimum truth serums with optional predictions. In *Proceedings of the 4th Workshop on Social Computing and User Generated Content (SC14)*.

Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., & Ramchandran, K. (2013). A case for ordinal peer-evaluation in moocs. In *NIPS Workshop on Data Driven Education*.

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (pp. 275–284). ACM.

Shnayder, V., Agarwal, A., Frongillo, R., & Parkes, D. C. (2016). Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (pp. 179–196). ACM.

Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*, 58–67.

Waggoner, B., & Chen, Y. (2014). Output agreement mechanisms and common knowledge. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing*.

Wilson, R. (1987). Game-theoretic approaches to trading processes. In *Advances in Economic Theory: Fifth World Congress* (pp. 33–77).

Witkowski, J., Bachrach, Y., Key, P., & 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*.

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

Witkowski, J., & Parkes, D. C. (2013). Learning the prior in minimal peer prediction. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content at the ACM Conference on Electronic Commerce* (p. 14). Citeseer.

Wright, J. R., Thornton, C., & Leyton-Brown, K. (2015). Mechanical TA: Partially automated high-stakes peer grading. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (pp. 96–101).

Zhang, P., & Chen, Y. (2014). Elicitability and knowledge-free elicitation with peer prediction. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multiagent Systems* (pp. 245–252). International Foundation for Autonomous Agents and Multiagent Systems.

Zohar, A., & Rosenschein, J. S. (2006). Robust mechanisms for information elicitation. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems* AAMAS '06 (pp. 1202–1204). New York, NY, USA: ACM. URL: `http://doi.acm.org/10.1145/1160633.1160849`. doi:10.1145/1160633.1160849.

Zohar, A., & Rosenschein, J. S. (2008). Mechanisms for information elicitation. *Artificial Intelligence*, *172*, 1917–1939.

Appendix A. Proof of Lemma 1

Lemma 1. *The minimum spot-check probability p_{ds} at which the truthful strategy is dominant for the peer-insensitive mechanism satisfies the following equation.*

$$p_{\mathrm{ds}}\,\mathbb{E}[y(s^h,s^t)] - c^E = p_{\mathrm{ds}}\,\mathbb{E}[y(g^l(s^l),s^t)]. \tag{A.1}$$

Proof. Consider the peer-insensitive mechanism with a fixed spot-check probability $p \geq 0$. When an agent uses the truthful strategy, his expected utility is

$$u_t = p\,\mathbb{E}[y(s^h,s^t)] + (1-p)\,W - c^E. \tag{A.2}$$

When an agent invests no effort, his best strategy is g^l. His expected utility from playing the g^l strategy is

$$u_l = p\,\mathbb{E}[y(g^l(s^l),s^t)] + (1-p)\,W. \tag{A.3}$$

All of $\mathbb{E}[y(s^h,s^t)]$, W, c^E, and $\mathbb{E}[y(g^l(s^l),s^t)]$ are constants. Therefore, u_t and u_l are both linear functions of p. When $p=0$, $u_t = W - c^E$ and $u_l = W$. Since $c^E > 0$, $u_t < u_l$ when $p = 0$. When the spot-check probability is 0, the agent prefers playing the best strategy conditional on investing no effort to playing the truthful strategy. When $p = 1$, $u_t = \mathbb{E}[y(s^h,s^t)] - c^E$, and $u_l = \mathbb{E}[y(g^l(s^l),s^t)])$. By our assumption (3), $u_t > u_l$. When the spot-check probability is 1, the agent prefers playing the truthful strategy to playing the best strategy conditional on investing no effort.

Since u_t and u_l are linear functions of p, there is a unique value of p in $[0,1]$ such that $u_t = u_l$. Denote this p value by p_{ds}. For any $p < p_{\mathrm{ds}}$, $u_l > u_t$ and an agent's expected utility for playing the best strategy conditional on investing no effort is higher than that of playing the truthful strategy. For any $p > p_{\mathrm{ds}}$, $u_l < u_t$ and an agent's expected utility for playing the truthful strategy is higher than that of playing the best strategy conditional on investing no effort.

When $p = p_{\mathrm{ds}}$, an agent's expected utilities by using the two strategies are the same. Thus, we solve for p_{ds} as follows.

$$\begin{gathered}
p_{\mathrm{ds}}\,\mathbb{E}[y(s^h,s^t)] + (1-p_{\mathrm{ds}})\,W - c^E = p_{\mathrm{ds}}\,\mathbb{E}[y(g^l(s^l),s^t)] + (1-p_{\mathrm{ds}})\,W \\
p_{\mathrm{ds}}\,\mathbb{E}[y(s^h,s^t)] - c^E = p_{\mathrm{ds}}\,\mathbb{E}[y(g^l(s^l),s^t)].
\end{gathered}$$

□

Appendix B. Proof of Lemma 2

Lemma 2. *For any spot-checking peer-prediction mechanism, if the g^l equilibrium exists when $c^E = 0$ and $p = 0$, then $p_{\mathrm{el}} \geq p_{\mathrm{ds}}$ for all $c^E \geq 0$.*

Proof. Consider a spot-checking peer prediction mechanism. Assume that, for this mechanism, the g^l equilibrium exists when $c^E = 0$ and $p = 0$.

First, we prove that p_{el} exists and is well defined. Recall that p_{el} is the minimum spot-check probability at which the g^l equilibrium is eliminated. We need to show that there is a unique spot-check probability p_{el} in $[0, 1]$ such that for all spot check probabilities less than or equal to p_{el}, the g^l equilibrium exists and for all spot check probabilities greater than p_{el}, the g^l equilibrium does not exist.

It suffices to focus on strategies conditional on investing full effort. By definition, the g^l strategy is the best strategy conditional on investing no effort. For any spot-check probability, a strategy conditional on investing no effort can never become a beneficial deviation to the g^l equilibrium.

First, we will show that, when $p = 0$, the g^l equilibrium exists for any $c^E \geq 0$. Let the spot-check probability be zero and consider any $c^E \geq 0$. Suppose that all agents except agent i play the g^l strategy. Agent i's expected utility for playing the g^l strategy is

$$\mathbb{E}[z(g^l(s^l), g^l(s^l))].$$

Agent i's expected utility for playing any strategy g conditional on full effort is:

$$\mathbb{E}[z(g(s^h), g^l(s^l))] - c^E.$$

We need to show that the first expression is greater than or equal to the second expression. By our assumptions, the following inequality is true.

$$\mathbb{E}[z(g^l(s^l), g^l(s^l))] \geq \mathbb{E}[z(g(s^h), g^l(s^l))] \geq \mathbb{E}[z(g(s^h), g^l(s^l))] - c^E,$$

where the first inequality was due to the fact that the g^l equilibrium exists when $c^E = 0$, and the second inequality was true because $c^E \geq 0$. Thus, when $p = 0$, an agent prefers to follow the g^l equilibrium than deviating to any other strategy.

Next, we show that, when $p = 1$, the g^l equilibrium does not exist. Let the spot-check probability be 1 and assume that all agents except agent i play the g^l strategy. Agent i's expected utility for playing the g^l strategy is

$$\mathbb{E}[y(g^l(s^l), s^t)].$$

Agent i's expected utility for playing the truthful strategy is

$$\mathbb{E}[y(s^h, s^t)] - c^E.$$

By assumption (3), the following inequality is true.

$$\mathbb{E}[y(s^h, s^t)] - c^E > \mathbb{E}[y(g^l(s^l), s^t)].$$

Therefore, when $p = 1$, the truthful strategy is a strictly beneficial deviation and the g^l equilibrium does not exist.

Next, we show that, p_{el} is a well defined threshold value. That is, for any spot-check probability less than or equal to p_{el}, the g^l equilibrium exists and for any spot-check probability greater than p_{el} the g^l equilibrium does not exist. Starting with a spot-check probability of zero, we increase the spot-check probability until there is a weakly beneficial deviation to the g^l equilibrium. We denote this spot-check probability by p_{el} and we call this weakly beneficial deviation $g^{\text{br}}(s^h)$. We will show that for any spot-check probability strictly greater than p_{el}, $g^{\text{br}}(s^h)$ is still a beneficial deviation to the g^l equilibrium.

Consider an arbitrary spot-check probability p. Suppose that all agents except agent i play the g^l strategy. If agent i plays the g^l strategy, his expected utility is

$$p\,\mathbb{E}[y(g^l(s^l), s^t)] + (1-p)\,\mathbb{E}[z(g^l(s^l), g^l(s^l))]. \tag{B.1}$$

If agent i plays the g^{br} strategy, his expected utility is

$$p\,\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] + (1-p)\,\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E. \tag{B.2}$$

p_{el} is the smallest spot-check probability at which there is a beneficial deviation to the g^l equilibrium. The two expressions above must be equal when the spot-check probability is p_{el}. Thus p_{el} must satisfy the following equation.

$$p_{\text{el}}\,\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] + (1-p_{\text{el}})\,\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E = p_{\text{el}}\,\mathbb{E}[y(g^l(s^l), s^t)] + (1-p_{\text{el}})\,\mathbb{E}[z(g^l(s^l), g^l(s^l))], \tag{B.3}$$
$$p_{\text{el}}\,\Big(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - \mathbb{E}[y(g^l(s^l), s^t)]\Big) + (1-p_{\text{el}})\,\Big(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l)) - \mathbb{E}[z(g^l(s^l), g^l(s^l))]]\Big) - c^E = 0.$$

Rewriting $-c^E$ to be $-p_{\text{el}}c^E - (1-p_{\text{el}})c^E$, the equation becomes:

$$p_{\text{el}}\,\big(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]]\big) + (1-p_{\text{el}})\,\big(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l)) - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]]\big)$$
$$= 0. \tag{B.4}$$

By our assumption, the g^l equilibrium exists when $p = 0$. Therefore, an agent's expected utility for deviating to $g^{\text{br}}(s^h)$ must be worse than following the g^l equilibrium, that is,

$$\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E \leq \mathbb{E}[z(g^l(s^l), g^l(s^l))],$$
$$\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))] \leq 0.$$

Thus, for equation (B.4) to hold, its first term must be non-negative. That is,

$$\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)] \geq 0.$$

We want to show that for any spot-check probability strictly greater than p_{el}, the $g^{\text{br}}(s^h)$ strategy is a beneficial deviation to the g^l equilibrium.

Consider a spot-check probability p that is strictly greater than p_{el}. Assume that all agents except agent i play the g^l strategy. Agent i's expected utility for playing the g^{br} strategy minus agent i's expected utility for playing the g^l strategy is

$$\begin{aligned}
&p\left(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E\right) + (1-p)\left(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E\right) \\
&\qquad - \big(\, p\,\mathbb{E}[y(g^l(s^l), s^t)] + (1-p)\,\mathbb{E}[z(g^l(s^l), g^l(s^l))]\,\big) \\
&= p\left(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]\right) + (1-p)\left(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\right).
\end{aligned}$$

Note that

$$p\left(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]\right) > p_{\text{el}}\left(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]\right)$$

because $\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]$ is greater than or equal to zero and $p > p_{\text{el}}$.

Similarly, note that

$$(1-p)\left(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\right) > (1-p_{\text{el}})\left(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\right)$$

because $\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]$ is less than or equal to zero and $1-p < 1-p_{\text{el}}$. Therefore,

$$\begin{aligned}
&p\left(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]\right) + (1-p)\left(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\right) \\
&> p_{\text{el}}\left(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]\right) + (1-p_{\text{el}})\left(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\right) \\
&= 0.
\end{aligned}$$

Rewriting the inequality, we have that

$$\begin{aligned}
&p\left(\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] - c^E - \mathbb{E}[y(g^l(s^l), s^t)]\right) + (1-p)\left(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\right) > 0, \\
&p\,\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] + (1-p)\,\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - c^E > p\,\mathbb{E}[y(g^l(s^l), s^t)] + (1-p)\,\mathbb{E}[z(g^l(s^l), g^l(s^l))]. \quad \text{(B.5)}
\end{aligned}$$

Therefore, for any spot-check probability $p > p_{\text{el}}$, the $g^{\text{br}}(s^h)$ strategy is a beneficial deviation to the g^l equilibrium. So the g^l equilibrium does not exist for any spot-check probability greater than p_{el}.

Next, we will show that $p_{\text{el}} \geq p_{\text{ds}}$.

Since the g^l equilibrium exists when $c^E = 0$ and $p = 0$, it follows from the definition of equilibrium that

$$\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] \leq \mathbb{E}[z(g^l(s^l), g^l(s^l))]. \quad \text{(B.6)}$$

Consider any $c^E \geq 0$. Taking p_{el} and substituting into the LHS of the definition of p_{ds} in equation (3), we can see that the LHS of the resulting equation (3) is greater than the RHS, as shown below.

$$\begin{aligned}
& p_{\text{el}}\,\mathbb{E}[y(s^h, s^t)] - c^E \\
& \geq p_{\text{el}}\,\mathbb{E}[y(s^h, s^t)] + (1 - p_{\text{el}})\,(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - \mathbb{E}[z(g^l(s^l), g^l(s^l))]) - c^E && \text{(B.7)} \\
& > p_{\text{el}}\,\mathbb{E}[y(g^{\text{br}}(s^h), s^t)] + (1 - p_{\text{el}})\,(\mathbb{E}[z(g^{\text{br}}(s^h), g^l(s^l))] - \mathbb{E}[z(g^l(s^l), g^l(s^l))]) - c^E && \text{(B.8)} \\
& = p_{\text{el}}\,\mathbb{E}[y(g^l(s^l), s^t)]. && \text{(B.9)}
\end{aligned}$$

Inequality (B.7) holds due to Equation (B.6). Inequality (B.8) holds because reporting the high-quality signal maximizes the spot-check reward. Equation (B.9) follows from Equation (B.3).

By definition of p_{ds}, it is the minimum spot-check probability for which the LHS of (3) is greater than or equal to its RHS. Thus, it must be that $p_{\text{el}} \geq p_{\text{ds}}$. □

Appendix C. Proof of Lemma 3

Lemma 3. *For any spot-checking peer-prediction mechanism, if the g^l equilibrium exists and Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$, then $p_{\text{ex}} \geq p_{\text{ds}}$ for all $c^E \geq 0$.*

Proof. Recall that p_{ex} is the minimum spot-check probability at which the g^l equilibrium Pareto dominates the truthful equilibrium while the g^l equilibrium exists at $p = p_{\text{ex}}$.

First, we prove that p_{ex} exists and is well defined, and we derive an expression for p_{ex}.

Consider a spot-checking peer prediction mechanism. For this mechanism, assume that the g^l equilibrium exists and Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$.

Consider a fixed spot-check probability $p \geq 0$. Assume that the g^l equilibrium exists at this spot-check probability. At the truthful equilibrium, an agent's expected utility is

$$u_t = p\,\mathbb{E}[y(s^h, s^t)] + (1 - p)\,\mathbb{E}[z(s^h, s^h)] - c^E. \tag{C.1}$$

At the g^l equilibrium, an agent's expected utility is

$$u_l = p\,\mathbb{E}[y(g^l(s^l), s^t)] + (1 - p)\,\mathbb{E}[z(g^l(s^l), g^l(s^l))]. \tag{C.2}$$

All of $\mathbb{E}[y(s^h, s^t)]$, $\mathbb{E}[z(s^h, s^h)]$, c^E, $\mathbb{E}[y(g^l(s^l), s^t)]$, and $\mathbb{E}[z(g^l(s^l), g^l(s^l))]$ are constants. Thus, u_t and u_l are both linear functions of p.

When the spot-check probability is 0, an agent's expected utilities at the truthful and the g^l equilibria are $u_t = \mathbb{E}[z(s^h, s^h)] - c^E$ and $u_l = \mathbb{E}[z(g^l(s^l), g^l(s^l))]$ respectively. Thus, we have that

$$\mathbb{E}[z(g^l(s^l), g^l(s^l))] \geq \mathbb{E}[z(s^h, s^h)] \geq \mathbb{E}[z(s^h, s^h)] - c^E.,$$

where the first inequality was due to our assumption that the g^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$, and the second inequality was due to $c^E \geq 0$. Therefore, when the spot-check probability is 0, an agent's expected utility at the g^l equilibrium is higher than his expected utility at the truthful equilibrium for any $c^E \geq 0$.

When the spot-check probability is 1, an agent's expected utilities at the truthful and the g^l equilibria are $u_t = \mathbb{E}[y(s^h, s^t)] - c^E$ and $u_l = \mathbb{E}[y(g^l(s^l), s^t)]$. We know that $\mathbb{E}[y(s^h, s^t)] - c^E > \mathbb{E}[y(g^l(s^l), s^t)]$ by by assumption (3). Thus, when the spot-check probability is 1, an agent's expected utility at the truthful equilibrium is higher than his expected utility at the g^l equilibrium.

Since u_t and u_l are linear functions of p, there is a unique value of p in $[0, 1]$ such that an agent's expected utilities at the two equilibria are the same. We denote this p value by p_{ex}. When $p < p_{\text{ex}}$, the agent's expected utility in the g^l equilibrium is higher. When $p > p_{\text{ex}}$, the agent's expected utility in the truthful equilibrium is higher. When $p = p_{\text{ex}}$, an agent has the same expected utility in both equilibria. Thus p_{ex} must satisfy

$$\begin{aligned}
&p_{\text{ex}}\,\mathbb{E}[y(s^h, s^t)] + (1 - p_{\text{ex}})\,\mathbb{E}[z(s^h, s^h)] - c^E \\
&\qquad = p_{\text{ex}}\,\mathbb{E}[y(g^l(s^l), s^t)] + (1 - p_{\text{ex}})\,\mathbb{E}[z(g^l(s^l), g^l(s^l))] \\
&p_{\text{ex}}\,\mathbb{E}[y(s^h, s^t)] + (1 - p_{\text{ex}})\,\big(\mathbb{E}[z(s^h, s^h)] - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\big) - c^E \\
&\qquad = p_{\text{ex}}\,\mathbb{E}[y(g^l(s^l), s^t)]. \qquad \text{(C.3)}
\end{aligned}$$

Next, we would like to show that $p_{\text{ex}} \geq p_{\text{ds}}$.

Since the g^l equilibrium exists and Pareto dominates the truthful equilibrium for $c^E = 0$ and $p = 0$, it follows from the definition of Pareto dominance that

$$\mathbb{E}[z(s^h, s^h)] \leq \mathbb{E}[z(g^l(s^l), g^l(s^l))]. \qquad \text{(C.4)}$$

Taking p_{ex} and substituting it into the LHS of Equation (3) (definition of p_{ds}), in a setting with arbitrary positive $c^E \geq 0$, we have

$$\begin{aligned}
&p_{\text{ex}}\,\mathbb{E}[y(s^h, s^t)] - c^E \\
&\geq p_{\text{ex}}\,\mathbb{E}[y(s^h, s^t)] + (1 - p_{\text{ex}})\,\big(\mathbb{E}[z(s^h, s^h)] - \mathbb{E}[z(g^l(s^l), g^l(s^l))]\big) - c^E \qquad \text{(C.5)} \\
&= p_{\text{ex}}\,\mathbb{E}[y(g^l(s^l), s^t)] \qquad \text{(C.6)}
\end{aligned}$$

Equation (C.5) follows from Equation (C.4). Equation (C.6) follows from Equation (C.3).

Thus, if we substitute p_{ex} into Equation (3), then the resulting LHS is weakly greater than the RHS. By definition of p_{ds}, it is the minimum spot-check probability for which the LHS of (3) is greater than its RHS. Thus, it must be that $p_{\text{ex}} \geq p_{\text{ds}}$. □

Appendix D. Proof of Theorem 2

Theorem 2 (Sufficient condition for Pareto comparison)**.** *For any spot-checking peer-prediction mechanism, if the g^l equilibrium exists and Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$, then $p_{\text{Pareto}} \geq p_{\text{ds}}$ for all $c^E \geq 0$.*

Proof. Consider any spot-checking peer prediction mechanism.

For the truthful equilibrium to be Pareto dominant, it is necessary that either the g^l equilibrium is eliminated or the truthful equilibrium Pareto dominates the g^l equilibrium while the g^l equilibrium exists. p_{el} is the minimum spot-check probability at which the g^l equilibrium is eliminated. p_{ex} is the minimum spot-check probability at which the truthful equilibrium Pareto dominates the g^l equilibrium while the g^l equilibrium exists at $p = p_{\text{ex}}$. Thus, the minimum of p_{el} and p_{ex} is a lower bound of p_{Pareto}. Formally

$$p_{\text{Pareto}} \geq \min(p_{\text{el}}, p_{\text{ex}}). \tag{D.1}$$

By assumption, the g^l equilibrium exists when $p = 0$. By Lemma 2, we have

$$p_{\text{el}} \geq p_{\text{ds}}. \tag{D.2}$$

By assumption, the g^l equilibrium exists and Pareto dominates the truthful equilibrium when $p = 0$. By Lemma 3, we have

$$p_{\text{ex}} \geq p_{\text{ds}}. \tag{D.3}$$

By Equations (D.1), (D.2) and (D.3), we have

$$\begin{aligned} p_{\text{Pareto}} &\geq \min(p_{\text{el}}, p_{\text{ex}}) \\ &\geq \min(p_{\text{ds}}, p_{\text{ex}}) \\ &\geq \min(p_{\text{ds}}, p_{\text{ds}}) \\ &= p_{\text{ds}}. \end{aligned}$$

□

Appendix E. Peer Prediction Mechanisms

Output-Agreement Mechanisms Output agreement mechanisms only collect signal reports from agents and reward an agent i for evaluating object j based on agents' signal reports for the object (Faltings et al., 2012; Witkowski et al., 2013; Waggoner & Chen, 2014).

The standard output agreement mechanism studied by Waggoner & Chen (2014) and Witkowski et al. (2013) gives agent i a constant reward exactly when agent i's signal report matches the signal report of another random agent i' evaluating the same object. Agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = \mathbb{1}_{r_{ij} = r_{i'j}}$$

where $\mathbb{1}$ is the indicator function.

The Faltings et al. (2012) mechanism also rewards the agents for agreement, but the amount of the reward is scaled by the empirical frequency of the signal report agreed upon. Agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = \alpha + \beta \frac{\mathbb{1}_{r_{ij}=r_{i'j}}}{F(r_{ij})}$$

where $\alpha > 0$ and $\beta > 0$ are constants and $F(r)$ is the empirical frequency of report r.

Multi-Object Mechanisms Multi-object mechanisms reward each agent based on agents' reports for multiple objects (Dasgupta & Ghosh, 2013; Radanovic & Faltings, 2015; Kamble et al., 2015; Shnayder et al., 2016).

The multi-signal Dasgupta-Ghosh mechanism (Dasgupta & Ghosh, 2013; Shnayder et al., 2016) and the Kamble et al. (2015) mechanism extend the output agreement mechanisms by adding additional scaling terms to the reward. These scaling terms are intended to exploit correlations between multiple tasks to make the truthful equilibrium Pareto dominate the uninformative equilibria, by reducing the reward to agents who agree to a report that is "unsurprising" given their reports on other objects.

The Kamble et al. (2015) mechanism adds a multiplicative scaling term to the reward for agreement, computed as follows. Choose two agents i' and i'' uniformly at random. For each signal $s \in Q$, let $f^j(s) = \mathbb{1}_{r_{i'j}=s}\mathbb{1}_{r_{i''j}=s}$. Define $\hat{f}(s) = \sqrt{\frac{1}{N}\sum_{j \in O} f^j(s)}$. Agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = \begin{cases} 0, \text{ if } \hat{f}(s) \in \{0, 1\} \\ \dfrac{K}{\hat{f}(r_{ij})}\mathbb{1}_{r_{ij}=r_{i'j}}, \text{ otherwise} \end{cases}$$

where K is a positive constant.

The multi-signal Dasgupta-Ghosh mechanism (Dasgupta & Ghosh, 2013; Shnayder et al., 2016) adds an additive scaling term to the reward for agreement, computed as follows. Suppose that agent i and i' both evaluate task j. Randomly choose two tasks j' and j'' such that agent i evaluated task j' but not task j'' and agent i' evaluated task j'' but not task j'. It is important that agents do not know which tasks are chosen to be j' and j''. Agent i is rewarded if his report matches that of agent i' on task j, and he is penalized if his report on object j' matches agent i' report on object j''. Formally, agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = \mathbb{1}_{r_{ij}=r_{i'j}} - \mathbb{1}_{r_{ij'}=r_{ij''}}.$$

The Radanovic & Faltings (2015) mechanism rewards the agents for report agreement using a reward function inspired by the quadratic scoring rule. The quadratic scoring rule is a proper scoring rule, which is designed to incentivize an agent to report his belief about the likelihoods of the outcomes of an event truthfully.

We can construct the Radanovic & Faltings (2015) mechanism's reward function as follows. Suppose that agent i evaluated object j. Choose another random agent i' who also evaluated object j. Construct a sample

Σ_i of reports which contains one report for every object that is not evaluated by agent i. The sample Σ_i is double-mixed if it contains all possible signal realizations at least twice. If Σ_i is not double-mixed, agent i's reward is 0. Otherwise, if Σ_i is double-mixed, the mechanism chooses two objects j' and j'' ($j' \neq j$, $j'' \neq j$ and $j' \neq j''$) such that the reports for objects j' and j'' in the sample are the same as agent i's report for j, i.e. $\Sigma_i(j') = \Sigma_i(j'') = r_{ij}$. For each of the objects j' and j'', randomly select two reports $r_{i''j'}$ and $r_{i'''j''}$. Agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = \frac{1}{2} + \mathbb{1}_{r_{i''j'}=r_{i'j}} - \frac{1}{2}\sum_{s \in Q} \mathbb{1}_{r_{i''j'}=s}\mathbb{1}_{r_{i'''j''}=s}.$$

Belief Based Mechanisms Belief based mechanisms collect both signal and belief reports from agents and reward each agent based on all agents' signal and belief reports for each object (Witkowski & Parkes, 2012, 2013; Radanovic & Faltings, 2013, 2014; Riley, 2014).

These mechanisms make use of proper scoring rules, which are designed to incentivize an agent to report his belief truthfully. Formally, a scoring rule is a function $R : \Delta(Q) \times Q \to \mathbb{R}$, which computes a real valued score based on a reported belief over the likelihoods of all possible signals and a realized signal. The scoring rule is proper if an agent's expected score is maximized when he reports his belief truthfully.

The robust Bayesian Truth Serum (BTS) (Witkowski & Parkes, 2012, 2013) rewards agent i for evaluating object j by how well his belief report b_{ij} and shadowed belief report b^s_{ij} predict the signal reports of another random agent i''. Agent i's shadowed belief report b^s_{ij} is the result of modifying another agent i''s belief report based on agent i's signal report. To calculate agent i's reward, randomly choose two other agents i' and i'' who evaluated object j. Agent i's shadowed belief report b^s_{ij} is calculated as follows. $b^s_{ij} = b_{i'j} + \delta$ if $r_{ij} = 1$ and $b^s_{ij} = b_{i'j} - \delta$ if $r_{ij} = 0$ where $\delta = \min(b_{i'j}, 1 - b_{i'j})$. Agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = R(b^s_{ij}, r_{i''j}) + R(b_{ij}, r_{i''j}).$$

The multi-valued robust BTS (Radanovic & Faltings, 2013) rewards agent i if his signal report matches the signal report of another random agent i' and his belief report accurately predicts agent i''s signal report. Formally, agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = \frac{1}{b_{i'j}(r_{ij})}\mathbb{1}_{r_{ij}=r_{i'j}} + R(b_{ij}, r_{i'j}).$$

The divergence-based BTS (Radanovic & Faltings, 2014) rewards agent i for evaluating object j if his belief report accurately predicts another random agent's signal report. Moreover, it penalizes agent i if his signal report matches the signal report of another agent i' but his belief report is sufficiently different from the belief report of agent i'. Formally, agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = -\mathbb{1}_{r_{ij}=r_{i'j}||D(b_{ij},b_{i'j})>\theta} + R(b_{ij}, r_{i'j})$$

where $D(||)$ is the divergence associated with the strictly proper scoring rule R, and θ is a parameter of the mechanism.

The Riley (2014) mechanism rewards agent i for evaluating object j by how well his belief report predicts other agents' signal reports. Moreover, agent i's reward is bounded above by the score for the average belief report of other agents who made the same signal report. Formally, let $\delta_i = \min_{s \in Q} |\{r_{i'} = s | i' \neq i\}|$ be the minimum number of other agents who have reported any given signal. Let $q_i(r_{ij})$ to be the average belief report for all other agents who made the same signal report as agent i. Agent i's reward is

$$z_{ij}(\mathbf{r}, \mathbf{b}) = \begin{cases} R(b_{ij}, r_{-ij}), \text{ if } \delta_i = 0 \\ \min\{R(b_{ij}, r_{-ij}), R(q_i(r_{ij}), r_{-ij})\}, \text{ if } \delta_i \geq 1. \end{cases}$$

Appendix F. Proof of Corollary 1

We begin with the following lemma.

Lemma 4. *For the spot-check mechanism in Equation* (2)*, conditional on investing no effort, an agent maximizes his spot-check reward by reporting the low-quality signal s^l truthfully.*

Proof. Consider the spot-check reward mechanism in equation (2).

If an agent invests no effort, he could either use a strategy that is independent of the low-quality signal or use a strategy that depends on the low-quality signal.

Case 1: Suppose that the agent uses a strategy that is independent of the low-quality signal. Assume that the agent uses a mixed strategy in which he reports r with probability p_r where $\sum_{r \in Q} p_r = 1$. Then the agent's expected utility is always zero, as shown below.

$$\sum_{r \in Q} p_r \Pr(s^t = r) - \sum_{r \in Q} p_r \Pr(s^t = r) = 0.$$

Case 2: Suppose that the agent uses a strategy where his report is a function of the low-quality signal. Assume that the agent uses a mixed strategy in which he reports r with probability p_{sr} when the realized low-quality signal is s. Then the agent's expected utility is shown below.

$$\sum_{s \in Q} \Pr(s^l = s) \left(\sum_{r \in Q} p_{sr} \Pr(s^t = r | s^l = s) \right) - \sum_{s \in Q} \Pr(s^l = s) \left(\sum_{r \in Q} p_{sr} \Pr(s^t = r) \right) \tag{F.1}$$

$$= \sum_{s, r \in Q} \big(\Pr(s^l = s, s^t = r) - \Pr(s^l = s) \Pr(s^t = r) \big) \tag{F.2}$$

Recall that the low-quality signal s^l is categorical with respect to the trusted report s^t. Thus, for any two realized signal $r, s \in Q$ where $\neq r$, we have that

$$\Pr(s^t = r | s^l = s) < \Pr(s^t = r) \tag{F.3}$$

$$\Pr(s^t = r | s^l = s) - \Pr(s^t = r) < 0 \tag{F.4}$$

$$\Pr(s^t = r, s^l = s) - \Pr(s^t = r) \Pr(s^l = s) < 0. \tag{F.5}$$

Note that $\sum_{r\in Q}\left(\Pr(s^t = r, s^l = s) - \Pr(s^t = r)\Pr(s^l = s)\right) = \Pr(s^l = s) - \Pr(s^l = s) = 0$. Thus, observing a realized signal must increase the probability that the trusted report is the same signal, i.e.

$$\Pr(s^t = s, s^l = s) - \Pr(s^t = s)\Pr(s^l = s), \tag{F.6}$$

for any realized signal $s \in Q$. In equation (F.2), any term with $s = r$ is positive and every other term is negative. Therefore, the agent's expected utility is maximized when $p_{ss} = 1, \forall s \in Q$, that is, the agent reports the low-quality signal truthfully. □

We can now prove Corollary 1:

Corollary 1. *For spot-checking peer-prediction mechanisms based on Faltings et al. (2012); Witkowski et al. (2013); Dasgupta & Ghosh (2013); Waggoner & Chen (2014); Kamble et al. (2015); Radanovic & Faltings (2015) and Shnayder et al. (2016), the minimum spot-check probability p_{Pareto} for the Pareto dominance of the truthful equilibrium is greater than or equal to the minimum spot-check probability p_{ds} at which the truthful strategy is a dominant strategy for the peer-insensitive mechanism.*

Proof. By Lemma 4, for any spot-checking peer prediction mechanism, the g^l strategy is to always report the low-quality signal s^l.

To verify that the conditions of Theorem 2 are satisfied, it suffices to verify that when $p = 0$, the s^l equilibrium of the peer prediction mechanism exists and Pareto dominates the truthful equilibrium. We verify these two conditions for all of the listed peer prediction mechanisms below.

We first consider output agreement peer prediction mechanisms.

The Standard Output Agreement Mechanism (Witkowski et al., 2013; Waggoner & Chen, 2014) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents except i report s^l, then agent i's best response is to also report s^l in order to perfectly agree with other reports.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, an agent's expected utility is

$$\sum_{s^l\in Q}\Pr(s^l)\Pr(s^l|s^l) = \sum_{s^l\in Q}\Pr(s^l) = 1,$$

where the equality is due to the fact that the low-quality signals are noiseless ($\Pr(s^l|s^l) = 1$).

When $c^E = 0$ and $p = 0$, at the truthful equilibrium, an agent's expected utility is

$$\sum_{s^h\in Q}\Pr(s^h)\Pr(s^h|s^h) < \sum_{s^h\in Q}\Pr(s^h) = 1,$$

where the inequality is due to the fact that the high-quality signals are noisy. That is, for every realization s^h of the high-quality signal, $\Pr(s^h|s^h) \leq 1$ and there exists one realization s^h of the high-quality signal such that $\Pr(s^h|s^h) < 1$. Thus, the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$. The conditions of Theorem 2 are satisfied.

Peer Truth Serum (Faltings et al., 2012) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents except i report s^l, then agent i's best response is to also report s^l.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, everyone reports s^l and the empirical frequency of s^l reports is 1 ($F(s^l) = 1$). Thus, every agent's expected utility is

$$\alpha + \beta \frac{1}{F(s^l)} = \alpha + \beta.$$

When $c^E = 0$ and $p = 0$, at the truthful equilibrium, if agent receives the high-quality signal s^h for an object, then he expects the empirical frequency of this signal to be $\Pr(s^h|s^h)$. Thus, at this equilibrium, an agent's expected utility is

$$\alpha + \beta \sum_{s^h \in Q} \Pr(s^h)\Pr(s^h|s^h) \frac{1}{\Pr(s^h|s^h)} = \alpha + \beta.$$

Thus, the s^l equilibrium (weakly) Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$. The conditions of Theorem 2 are satisfied.

Next, we consider multi-object peer prediction mechanisms.

The Multi-Signal Dasgupta-Ghosh Mechanism (Dasgupta & Ghosh, 2013; Shnayder et al., 2016) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents always report the low-quality signal s^l for every object, then agent i's best response is also to report s^l in order to maximize the probability of his report agreeing with other agents' reports for the same object.)

When $p = 0$, at the s^l equilibrium, an agent's expected utility is

$$\begin{aligned} &\sum_{s^l \in Q} \Pr(s^l)\Pr(s^l|s^l) - \sum_{s^l \in Q} \Pr(s^l)\Pr(s^l) = \sum_{s^l \in Q} \Pr(s^l) - \sum_{s^l \in Q} \Pr(s^l)\Pr(s^l) \\ &= 1 - \sum_{s^l \in Q} \Pr(s^l)^2 = 1 - \sum_{s^l \in Q} \frac{1}{|Q|^2} = 1 - \frac{1}{|Q|}, \end{aligned}$$

where the first equality was due to the fact that the low-quality signal s^l is noiseless ($\Pr(s^l|s^l) = 1$) and the second equality was due to the fact that s^l is drawn from a uniform distribution ($\Pr(s^l) = \frac{1}{|Q|}$).

When $c^E = 0$ and $p = 0$, at the truthful equilibrium, an agent's expected utility is

$$\begin{aligned} &\sum_{s^h \in Q} \Pr(s^h)\Pr(s^h|s^h) - \sum_{s^h \in Q} \Pr(s^h)\Pr(s^h) < \sum_{s^h \in Q} \Pr(s^h) - \sum_{s^h \in Q} \Pr(s^h)^2 \\ &= 1 - \sum_{s^h \in Q} \Pr(s^h)^2 \leq 1 - \frac{1}{|Q|}, \end{aligned}$$

where the first inequality was due to the fact that the high-quality signal is noisy. That is, for every realization s^h of the high-quality signal, $\Pr(s^h|s^h) \leq 1$ and there exists one realization s^h of the high-quality signal such that $\Pr(s^h|s^h) < 1$. Thus, the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$. The conditions of Theorem 2 are satisfied.

Kamble et al. (2015) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents always report s^l, an agent's best response is also to report s^l because doing so maximizes the probability of his report agreeing with other agents' reports for the same object.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, an agent's expected utility is

$$\begin{aligned}&\sum_{s^l \in Q} \Pr(s^l)\Pr(s^l|s^l) \lim_{N\to\infty} r(s^l) = \sum_{s^l \in Q} \Pr(s^l)\frac{K}{\sqrt{\Pr(s^l, s^l)}} = K\sum_{s^l \in Q}\frac{\Pr(s^l)}{\sqrt{\Pr(s^l)}}\\ &= K\sum_{s^l \in Q}\sqrt{\Pr(s^l)} = K\sum_{s_l \in Q}\sqrt{\frac{1}{|Q|}},\end{aligned}$$

where the first two equalities were due to the fact that the low-quality signal s^l is noiseless ($\Pr(s^l|s^l) = \Pr(s^l)$), and the final equality was due to the fact that the low-quality signal s^l is drawn from a uniform distribution.

When $c^E = 0$ and $p = 0$, at the truthful equilibrium, an agent's expected utility is

$$\begin{aligned}&\sum_{s^h \in Q} \Pr(s^h)\Pr(s^h|s^h) \lim_{N\to\infty} r(s^h) = \sum_{s^h \in Q} \Pr(s^h, s^h)\frac{K}{\sqrt{\Pr(s^h, s^h)}}\\ &= K\sum_{s^h \in Q}\sqrt{\Pr(s^h, s^h)} < K\sum_{s^h \in Q}\sqrt{\Pr(s^h)} \leq K\sum_{s^h \in Q}\sqrt{\frac{1}{|Q|}},\end{aligned}$$

where the first inequality was due to the fact that the high-quality signal s^h is noisy. That is, for every realization s^h of the high-quality signal, $\Pr(s^h|s^h) \leq 1$ and there exists one realization s^h of the high-quality signal such that $\Pr(s^h|s^h) < 1$. Thus, the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$. The conditions of Theorem 2 are satisfied.

Radanovic & Faltings (2015) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents always report s^l for every object, then any sample taken will not be "double mixed".[6] Thus, an agent's expected utility is zero regardless of his strategy. In particular also reporting s^l for every object is a best response.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, it must be that $r_{i''j'} = r_{i'j}$ and $r_{i''j'} = r_{i'''j''} = r_{ij}$. An agent's expected utility at the s^l equilibrium is:

$$\frac{1}{2} + \mathbb{1}_{r_{i''j'} = r_{i'j}} - \frac{1}{2}\sum_{s \in Q}\mathbb{1}_{r_{i''j'} = s}\mathbb{1}_{r_{i'''j''} = s} = \frac{1}{2} + 1 - \frac{1}{2} * 1 = 1.$$

Let $\pi(\Sigma)$ be the probability that the sample Σ is double mixed. When $c^E = 0$ and $p = 0$, at the truthful

[6] A sample is double mixed if every possible value appears at least twice. This mechanism behaves differently depending on whether or not it collects a double mixed sample of reports from the agents.

equilibrium, an agent's expected utility is:

$$\pi(\Sigma)\left(\frac{1}{2}+\Pr(r_{i''j'}|r_{ij})-\frac{1}{2}\sum_{s\in Q}\Pr(s|r_{ij})^2\right)\leq\frac{1}{2}+\Pr(r_{i''j'}|r_{ij})-\frac{1}{2}\sum_{s\in Q}\Pr(s|r_{ij})^2$$
$$\leq\frac{1}{2}+1-\frac{1}{2}*1=1,$$

where the first inequality is due to the fact that $\pi(\Sigma)\leq 1$ and the second inequality was due to the fact that the agent's expected utility is maximized when $\Pr(r_{i''j'}|r_{ij})=1$. Thus, the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E=0$ and $p=0$. The conditions of Theorem 2 are satisfied. □

Appendix G. Proof of Corollary 2

Corollary 2. *For spot-checking peer-prediction mechanisms based on Witkowski & Parkes (2012, 2013); Radanovic & Faltings (2013, 2014) and Riley (2014), if the peer-prediction mechanism uses a symmetric proper scoring rule, then the minimum spot-check probability p_{Pareto} for the Pareto dominance of the truthful equilibrium is greater than or equal to the minimum spot-check probability p_{ds} at which the truthful strategy is a dominant strategy for the peer-insensitive mechanism.*

Proof. By Lemma 4, for any spot-checking peer prediction mechanism, the g^l strategy is to always report the low-quality signal s^l.

To verify that the conditions of Theorem 2 are satisfied, it suffices to verify that when $p=0$, the s^l equilibrium of the peer prediction mechanism exists and Pareto dominates the truthful equilibrium. We verify these two conditions for all of the listed peer prediction mechanisms below.

Let b_s denote a belief report which predicts that signal s is observed with probability 1, i.e. $\Pr(s)=1$ and $\Pr(s')=0,\forall s'\in Q, s'\neq s$. Let the s^l equilibrium denote the equilibrium where every agent's signal report is s^l and belief report is b_{s^l}.

For mathematical convenience, we assume that the scoring rule is *symmetric* (Gneiting & Raftery, 2007). That is, the reward for reporting a signal that is predicted with probability 1 is the same regardless of the signal's identity:

$$R(b_s,s)=R(b_{s'},s'),\forall s\neq s'.$$

This is a very mild condition that is satisfied by all standard scoring rules that compute rewards based purely on the predicted probabilities and the outcome, including the quadratic scoring rule and the log scoring rule.

For symmetric scoring rules, when $p=0$, an agent's expected score is maximized by predicting b_s when s is observed for any signal $s\in Q$.

Binary Robust BTS (Witkowski & Parkes, 2012, 2013) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents report s^l and b_{s^l}, then the best belief report for agent i is b_{s^l}. Moreover the best signal report for agent i is s^l which leads to a shadowed belief report of b_{s^l}.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, an agent's expected utility is $R(b_{s^l}, s^l) + R(s_{s^l}, s^l)$. This is the maximum possible expected utility that an agent can achieve because the proper scoring rule R is symmetric. Therefore, it must be greater than or equal to the agent's expected utility at the truthful equilibrium when $c^E = 0$ and $p = 0$.

Multi-valued Robust BTS (Radanovic & Faltings, 2013) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents report s^l and b_{s^l}, then the best belief report for agent i is b_{s^l}. Moreover, the best signal report for agent i is s^l which maximizes the probability of his signal report agreeing with other agents' signal reports.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, an agent's expected utility is

$$\sum_{s^l} \Pr(s^l)\Pr(s^l|s^l) + R(b_{s^l}, s^l) = \sum_{s^l} \Pr(s^l) + R(b_{s^l}, s^l) = 1 + R(b_{s^l}, s^l),$$

where the first equality was due to the fact that the low-quality signal s^l is noiseless ($\Pr(s^l|s^l) = 1$).

When $c^E = 0$ and $p = 0$, at the truthful equilibrium, an agent's expected utility is

$$\begin{aligned}&\sum_{s^h \in Q} \Pr(s^h)\Pr(s^h|s^h)\frac{1}{\Pr(s^h|s^h)} + \mathbb{E}[R(\Pr(r_j|s^h), r_j)]\\ &= \sum_{s^h \in Q} \Pr(s^h) + \mathbb{E}[R(\Pr(r_j|s^h), r_j)] = 1 + \mathbb{E}[R(\Pr(r_j|s^h), r_j)] \leq 1 + R(b_{s^l}, s^l),\end{aligned}$$

where the inequality was due to the fact that the proper scoring rule R is symmetric. Thus, the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$. The conditions of Theorem 2 are therefore satisfied, and hence $p_{\text{Pareto}} \geq p_{\text{ds}}$ for all settings with positive effort cost $c^E \geq 0$.

Divergence-Based BTS (Radanovic & Faltings, 2014) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (If all other agents report s^l and b_{s^l}, then the best belief report for agent i is b_{s^l}. Moreover, the best signal report for agent i is s^l, which means that the penalty is 0 because the agent's signal reports agree and their belief reports also agree.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, an agent's expected utility is

$$-\mathbb{1}_{s^l = s^l || D(b_{s^l}, b_{s^l}) > \theta} + R(b_{s^l}, s^l) = R(b_{s^l}, s^l).$$

At the truthful equilibrium, an agent's expected utility is

$$-\mathbb{1}_{s^h_{i'j} = s^h_{i'j} || D(\Pr(r|s^h_{ij}), \Pr(r|s^h_{i'j})) > \theta} + R(\Pr(r|s^h), s^h) < R(\Pr(r|s^h), s^h) < R(b_{s^l}, s^l),$$

where the first inequality was due to the fact that the high-quality signal s^l is noisy. That is, for every realization s^h of the high-quality signal, $\Pr(s^h|s^h) \leq 1$ and there exists one realization s^h of the high-quality signal such that $\Pr(s^h|s^h) < 1$. The second inequality was due to the fact that the proper scoring rule R is symmetric. Thus, the s^l equilibrium Pareto dominates the truthful equilibrium when $c^E = 0$ and $p = 0$. The conditions of Theorem 2 are therefore satisfied, and hence $p_{\text{Pareto}} \geq p_{\text{ds}}$ for all settings with positive effort cost $c^E \geq 0$.

Riley (2014) When $c^E = 0$ and $p = 0$, the s^l equilibrium exists. (When all other agents always report s^l, for agent i, $\delta_i = 0$ because for any signal other than s^l, the number of other agents who reported the signal is 0. Thus, agent i's reward is $R(b_i, s^l)$. Since agent i's signal report does not affect his reward, reporting s^l is as good as reporting any other value. Moreover, since all other agents report s^l, the best belief report for agent i is to report b_{s^l}.)

When $c^E = 0$ and $p = 0$, at the s^l equilibrium, $\delta_i = 0$ because for any signal other than s^l, the number of other agents who reported the signal is 0. Thus, an agent's expected utility is $R(b_{s^l}, s^l)$. By the definition of the mechanism, an agent's reward is at most $R(b_i, r_{-i})$, which is less than or equal to $R(b_{s^l}, s^l)$ because R is a symmetric proper scoring rule. Therefore, an agent achieves the maximum expected utility at the s^l equilibrium, which is greater than or equal to the agent's expected utility at the truthful equilibrium when $c^E = 0$ and $p = 0$. □