Equilibrium Selection in Information Elicitation without Verification via Information Monotonicity*

Yuqing Kong[†1] **and Grant Schoenebeck**[‡2]

1 **University of Michigan, Ann Arbor, USA**
`yuqkong@umich.edu`
2 **University of Michigan, Ann Arbor, USA**
`schoeneb@umich.edu`

Abstract

In this paper, we propose a new mechanism – the Disagreement Mechanism – which elicits privately-held, non-variable information from self-interested agents in the single question (peer-prediction) setting.

To the best of our knowledge, our Disagreement Mechanism is the first strictly truthful mechanism in the single-question setting that is simultaneously:

- **Detail-Free:** does not need to know the common prior;
- **Focal:** truth-telling pays strictly higher than any other symmetric equilibria excluding some unnatural permutation equilibria;
- **Small group:** the properties of the mechanism hold even for a small number of agents,

even in binary signal setting. Our mechanism only asks each agent her signal as well as a forecast of the other agents' signals.

Additionally, we show that the focal result is both tight and robust, and we extend it to the case of asymmetric equilibria when the number of agents is sufficiently large.

1998 ACM Subject Classification J.4 Social and Behavioral Sciences

Keywords and phrases peer prediction, equilibrium selection, information theory

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.13

1 Introduction

User feedback requests (e.g Ebay's reputation system and the innumerable survey requests in one's email inbox) are increasingly prominent and important. However, the overwhelming number of requests can lead to low participation rates, which in turn may yield unrepresentative samples. To encourage participation, a system can reward people for answering requests. But this may cause perverse incentives: some people may answer a large of number of questions simply for the reward and without making any attempt to answer accurately. In this case, the reviews the system obtains may be inaccurate and meaningless. Moreover, people may be motivated to lie when they face a potential loss of privacy or can benefit in the future by lying now.

It is thus important to develop systems that motivate honesty. If we can verify the information people provide in the future (e.g via prediction markets), we can motivate

* A full version of the paper is available at [15], `https://arxiv.org/abs/1603.07751`

† The author was supported by National Science Foundation Career Award 1452915 and CCF Award 1618187.

‡ The author was supported by National Science Foundation Career Award 1452915 and Algorithms in the Field Award 1535912.

9th Innovations in Theoretical Computer Science Conference (ITCS 2018).
Editor: Anna R. Karlin; Article No. 13; pp. 13:1–13:20
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

honesty via this future verification. However, sometimes we need to elicit information without verification since the objective truth is hard to access or even does not exist (e.g. a self-report survey for involvement in crime). In our paper, we focus on the situation where the objective truth is not observable.

One important framework for designing incentive systems without verification is *peer prediction* [19]. *Peer prediction* uses each person's information to predict other people's information and pays according to how good the prediction is. *Peer prediction* assumes people's information is related and the systems and the people share a common prior. In the peer prediction mechanism, if an agent believes everyone else tells the truth, the best strategy to maximize her expected payment is telling the truth as well. In other words, peer prediction is *truthful* in the sense it has truth-telling as an equilibrium.

A series of work [20, 27, 21, 22, 30, 24, 7, 28, 29, 26, 11, 12, 16] extends peer prediction to incorporate some other desired properties in addition to *truthful* property, which we highlight here.

Detail-Free: A detail free mechanism does not need to know the common prior which is a clear advantage if deploying these mechanisms in the real world.

Focal: We hope the mechanism can pay the truth-telling equilibrium strictly more than other equilibria in expectation. However, we show this is not technically possible in the detail-free setting, and so instead we say *a mechanism is focal if truth-telling pays strictly higher than any other symmetric equilibria excluding some unnatural permutation equilibria* (see definition for permutation equilibria in Section 2). We emphasize that the *focal* property is very important, since in a non-focal mechanism, non-informative and effortless equilibria like everyone reporting the a priori most likely answer may pay equally or even much better than the truth-telling equilibrium, possibly incentivizing agents to coordinate on effortless and non-informative equilibria. Recent research [9] indicates that individuals in lab experiments do not always truth-tell when faced with peer prediction mechanisms; this may in part be related to the issue of equilibrium multiplicity.

Small group: The properties of the mechanism hold even when the number of agents is small.

General Informative Symmetric Common Prior: The mechanism should make minimal assumptions on the common prior. It is required that the prior be "informative" (that is each agents' signals contain stochastic information about the other agent's signals).

Finite Signals: The mechanism should only assume that the number of signals is finite, not binary.

Another desirable mechanism design property is the minimal property. A mechanism is minimal if it only requires agents to report their information rather than forecasts for other agents' reports. However, no detail free, strictly truthful mechanism can be minimal[1] [21]. We will introduce several strictly truthful, minimal mechanisms (not detail free) in the related work. Our mechanism is not minimal since it is detail free and strictly truthful.

Our Contributions

To the best of our knowledge, our Disagreement Mechanism is the first mechanism in the single-question setting that simultaneously has the *small group*, *focal*, *strictly truthful* and

[1] There exists detail free, strictly truthful mechanism that is "almost" minimal. For example, Riley [24] designs strictly truthful, detail free mechanism with optional prediction report.

Table 1 Detail Free Multi-Signal Single-Question Mechanisms Comparison

	Strictly Truthful	Small group	Focal	General Symmetric Prior
Bayesian Truth Serum [20]	$\times$		$\times$	$\leftthreetimes$[a]
Logarithmic PTS [23]	$\times$		$\times$	$\leftthreetimes$[a]
Multi-Signal SM [26]	$\times$	$\times$		
Multi-Valued RBTS [21]	$\times$	$\times$		
Minimal Truth Serum [24]	$\times$	$\times$		$\times$
Divergence-Based BTS [22]	$\times$	$\times$		$\times$
Disagreement Mechanism	$\times$	$\times$	$\times$	$\times$

[a] BTS requires an additional assumption – conditioning on the state, the signals are independently assigned to agents. But this conditional independence assumption is very natural in this literature, thus we still put $\leftthreetimes$ here.

detail free properties (see Table 1), even in binary signal setting. Moreover, our Disagreement Mechanism can be applied to a general family of symmetric prior, and

1. we show the *focal* property is **tight** in the sense that for the set of symmetric equilibria (what we call permutation equilibria) that have the same expected payment with truth-telling in our Disagreement Mechanism, no detail free, truthful mechanism can pay truth-telling strictly higher than permutation equilibria.
2. we show the *focal* property is **robust** in the sense that any symmetric equilibrium that has expected payoff close to truth-telling must be "close" to a permutation equilibrium;
3. we extend this *focal* property to asymmetric equilibria when the number of agents is sufficiently large in the sense that any asymmetric equilibrium which is "close" to a permutation equilibrium has expected payoff "close" to that of the truth-telling equilibrium, and any equilibrium that is not "close" to a permuation equilibrium pays strictly less then the truth-telling equilibrium.

The permutation equilibria are intuitively unnatural and risky as they require extreme coordination amongst the agents and as much effort as truth-telling. For *symmetric* equilibria that are not permutation equilibrium, **every agent's** expected payment is strictly less than the expected payment she obtains when everyone tells the truth. Thus our results about symmetric equilibrium are quite strong, despite the impossibility result. Asymmetric equilibrium require more coordination between agents than symmetric equilibrium. Additionally, the possible *total* gains from doing so are limited and go to zero as the number of agents increase.

High Level Techniques

Our *Disagreement Mechanism* pays agents individually (locally) for "agreement" and globally for "disagreement". When agents collude, they share information about their strategies. Since they are paid individually for "agreement", they will use their information about other people's strategies and "agree to agree" which reduces their global payoff which depends on "disagreement".

Essentially, when agents collude, our *Disagreement Mechanism* encourages each agent to implicitly admit their collusion by unilaterally increasing their individual payoff for doing so, but the mechanism then simultaneously decreases the total payoffs to all agents. Only when agents do not choose to collude and lack information about other people's strategies, can they “agree to disagree.” In this case, even when they maximize their individual payoff, globally they still have a lot of disagreement.

Technical Contributions

In addition to the above results, our works has several contributions in the techniques employed:

1. Our *Disagreement Mechanism* encourages not only agents with the same private information to agree, but also agents with different private information to disagree. We present a novel way to measure the amount of “information” by casting the reports as multiple labelled points in a space and measuring the quality of the “classification”.
2. To show that the “classification” quality always decreases with non-truthful equilibria, we exploit tools from information theory, namely ***Information Monotonicity***. Despite their natural and powerful application, to our knowledge, this is the first time such tools have been explicitly employed in the peer prediction literature.

1.1 Related Work

After Miller et al. [19] introducing peer prediction, a host of results (see, e.g., [20, 27, 21, 22, 30, 24, 7, 28, 29, 26, 11, 12, 16]) have followed. In this section, we will introduce them and classifies them into several categories according to the properties they (do not) have.

(1) Single-question, detail free, focal (not small group): Bayesian Truth Serum (BTS) [20] first successfully weakened the known common prior assumption (detail free) and solves the equilibrium multiplicity issue (focal). Prelec [20] also provides an important framework for mechanisms without known common prior. BTS requires the agents report – in addition to their reported signal – a forecast (prediction) of the other agents' reported signals, and uses this predictions in lieu of the common prior. BTS incentives agents to report accurate forecasts by rewarding forecasts that have the ability to predict the other agents' reported signal. However, BTS has two weakness: (1) BTS requires that the number of agents goes to infinity (or is large enough in a modified version) since the mechanism needs agents to believe it has access to the true distribution from which agents' signals are drawn. (2) The analysis of non-truthful equilibria provided in [20] requires that the number of agents goes to infinity and only proves that truth-telling has total expected payment at least as high as other equilibrium. Specifically, it does not rule out the existence of many other equilibrium which are all paid the same as the truth-telling equilibrium.

Logarithmic Peer Truth Serum (PTS) [23] extends BTS to a slightly different setting involving sensors, but still requires a large number of agents.

(2) Single-question, small group, detail free (not focal): Several mechanisms [27, 21, 22, 24, 28, 29, 26] are based on the BTS framework and address the first weakness of BTS. *Robust Bayesian Truth Serum (RBTS)* [27] is a mechanism which can only be applied to binary signals. *Multi-Valued RBTS* [21] and *Multi-Signal Shadowing Method (Multi-Signal SM)* [26] can be applied to non-binary signals while they require an *additional assumption* that an agent will think the probability that other agents receive signal σ higher if he himself also receives σ. *Divergence-based BTS* [22] can be applied to non-binary signals and does not require additional assumptions on the prior. All of those works do not solve the equilibrium multiplicity issue, but do work for a small number of agents. *Minimal Truth Serum (MTS)* [24] is a mechanism where agents have the option to report or not report their predictions, and also lacks analysis of non-truthful equilibria. MTS uses a typical zero-sum technique such that all equilibria are paid equally. In contrast, we show that

in our *Disagreement Mechanism* any equilibrium that is even close to paying more than the truth-telling equilibrium must be close to a small set of permutation equilibrium. The *Divergence based BTS* only requires the common prior assumption to be truthful. Because of its generality, we use it as a building block in our Disagreement Mechanism. However, the *Divergence based BTS* contains effortless equilibrium that pay significantly more than truth-telling. Moreover, analyzing the set of equilibria in *Divergence-based BTS* is very complicated and becomes a main technical obstacle in our paper. Thus, while the above work addresses the first weakness of BTS, it exacerbates the second.

(3) Single-question, small group, minimal (not detail free): Jurca and Faltings [11, 12] use algorithmic mechanism design to build their own peer prediction style mechanism where truth-telling is paid strictly better than non-truthful pure strategies but leave the analysis of mixed strategies as a open question. Frongillo and Witkowski [8] consider the design for robust, truthful and minimal peer prediction mechanisms with the prior knowledge and lack the analysis of non-truthful equilibria. Kong et al. [16] modify the peer prediction mechanism such that truth-telling is paid strictly better than any other non-truthful equilibrium. Thus, the mechanism designed in Kong et al. [16] is focal. Additionally, they optimize the cost their mechanism needs over a natural space. The assumption that the mechanism knows the prior, allows these mechanisms to not ask for a prediction report. However, unlike the current work, the mechanism still needs to know the prior and the analysis only works for the case of binary signals.

(4) Different Settings: Several works are not in the traditional one-question setting. Some of these works make the additional assumption of multiple a priori similar questions so that the mechanism need not explicitly ask the agents for a prediction. The mechanism in Dasgupta and Ghosh [6] uses the presence of multiple questions to elicit agent strategies with high effort, addressing the equilibrium multiplicity issue for binary signals in their setting. Recently Kong and Schoenebeck [14] and Shnayder et al. [25] independently extend the mechanism in [6] to non-binary signal setting but still require the presence of multiple questions. Our setting is different since the agents only have one question (and thus we do not have to assume relations between questions) and our results for equilibrium multiplicity issue are robust to non-binary signals.

In addition to the multiple questions setting, there are many other works in the settings that are different from our results. For example, Cai et al. [4] and Liu and Chen [17] consider the machine learning setting. Kamble et al. [13], Mandal et al. [18], and Agarwal et al. [1] consider the heterogeneous setting in the multiple questions setting. Zhang and Chen [30] consider a sequential game. Faltings et al. [7] consider a setting where they have an estimation of the public distribution of previous answers on other a priori similar questions.

2 Preliminaries, Background, and Notation

We recommend that the eager reader skip to the end of Section 2.4 where Hellinger Divergence is discussed and then refer back to the earlier preliminaries only as needed. Section 3 states the main theorem and outlines the technical contributions.

See notation table in Appendix A.

2.1 Prior Definitions and Assumptions

We consider a setting with n agents and a set of signals Σ, and define a *setting* as a tuple (n, Σ). Each agent i has a private signal $\sigma_i \in \Sigma$ chosen from a joint distribution Q over Σ^n called the prior. Given a prior Q, for $\sigma \in \Sigma$, let $q_i(\sigma) = \Pr_Q[\sigma_i = \sigma]$ be the *a priori* probability that agent i receives signal σ. Let $q_{j,i}(\sigma'|\sigma) = \Pr_Q[\sigma_j = \sigma'|\sigma_i = \sigma]$ be the probability that agent j receives signal σ given that agent i received signal σ'.

We say that a prior Q over Σ is *symmetric* if for all σ, $\sigma' \in \Sigma$ and for all pairs of agents $i \neq j$ and $i' \neq j'$ we have $q_i(\sigma) = q_{i'}(\sigma)$ and $q_{i,j}(\sigma|\sigma') = q_{i',j'}(\sigma|\sigma')$. That is, the first two moments of the prior do not depend on the agent identities. Because we will assume that the prior is symmetric, we denote $q_i(\sigma)$ by $q(\sigma)$ and $q_{i,j}(\sigma|\sigma')$ (where $i \neq j$) by $q(\sigma|\sigma')$. We also define $\mathbf{q}_\sigma = q(\cdot|\sigma)$. We assume the common prior shared by agents is

symmetric: we assume throughout that the agents' signals $\boldsymbol{\sigma}$ are drawn from some joint **symmetric prior** Q;

non-zero: for any $\sigma, \sigma' \in \Sigma$, $q(\sigma) > 0, q(\sigma|\sigma') > 0$;

informative: we assume if agents have different private signals, they will have different expectations for the fraction of at least one signal. That is for any $\sigma \neq \sigma'$, there exists σ'' such that $q(\sigma''|\sigma) \neq q(\sigma''|\sigma')$;

fine-grained: this assumption conceptually states that one state is not just a more likely version of another state. While we defer the exact definition to the full version, a slightly stronger version of this assumption is that $q(\sigma|\cdot)$ are linearly independent. This additional fine-grained assumption is required only when we need to show truth-telling is *strictly* better than other symmetric equilibria that are not permutation equilibria;

ensemble: the first two moments of the prior are fixed as the number of agents increases,

We sometimes will denote the class of priors that satisfy all five of these assumptions as SNIFE priors (see formal definitions of the above SNIFE prior assumptions in Appendix B.).

2.2 Game Setting and Equilibrium Concepts

Given a setting (n, Σ) with prior Q, we consider a game in which each agent i is asked to report his private signal $\sigma_i \in \Sigma$ and his prediction $\mathbf{p}_i \in \Delta_\Sigma$, a distribution over Σ, where $\mathbf{p}_i = \mathbf{q}_{\sigma_i}$. For any $\sigma \in \Sigma$, $\mathbf{p_i}(\sigma)$ is agent i's (reported) expectation for the fraction of other agents who has received σ given he has received σ_i. However, agents may not tell the truth. We denote the spaces of reports as $\Sigma \times \Delta_\Sigma$ by $\mathcal{R}$. We define a report profile of agent i as $r_i = (\hat{\sigma_i}, \hat{\mathbf{p}}_i) \in \mathcal{R}$ where $\hat{\sigma_i}$ is agent i's reported signal and $\hat{\mathbf{p}}_i$ is agent i's reported prediction.

We would like to encourage **truth-telling**, namely that agent i reports $\hat{\sigma_i} = \sigma_i, \hat{\mathbf{p}}_i = \mathbf{q}_{\sigma_i}$. To this end, agent i will receive some payment $\nu_i(\hat{\sigma}_i, \hat{\mathbf{p}}_i, \hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i})$ from our mechanism where $(\hat{\sigma}_{-i}, \hat{\mathbf{p}}_{-i})$ are all agents' report profiles excluding agent i.

We define **strategy** as a mapping from each possible signal σ and prior Q to a distribution over the report profile space.

We define a **strategy profile s** as a profile of all agents' strategies $\{s_1, s_2, ...s_n\}$ and we say agents play $\mathbf{s}$ if for any i, agent i plays strategy s_i. We say a strategy profile is *symmetric* if each agent plays the same strategy. Assuming a fixed prior Q, for any strategy profile $\boldsymbol{s} = (s_1, s_2, ..., s_n)$, we will represent the marginal distribution of an agent i's strategy for her signal report as a matrix θ_i where $\theta_i(\hat{\sigma}, \sigma)$ is the probability that agent will report signal $\hat{\sigma}$ when his private signal is σ. Note that θ_i is a *transition matrix*, since its entries are all non-negative and the sum of its every column is 1. We call θ_i the **signal strategy** of agent i. We also call $(\theta_1, \theta_2, ..., \theta_n)$ the signal strategy of $\boldsymbol{s}$. In the symmetric case, we call θ the signal strategy of $\boldsymbol{s}$. We say a signal strategy θ is **τ-close** to a permutation matrix if for any row of θ, there is at most one entry that is greater than τ.

We informally define a **permutation strategy profile** as a strategy profile where agents "collude" to relabel the signals and then tell the truth with respect to the relabelled signals. When agents play a permutation strategy profile, they play the same signal strategy which is a permutation matrix θ_π. We defer the formal definition of the permutation strategy profile to the full version. We will show that if truth-telling is an equilibrium, then all permutation strategy profiles are equilibria as well. Thus we sometimes call these permutation strategy profiles **permutation equilibria**. We define the **agent welfare** of a strategy profile $\mathbf{s}$ and a mechanism $\mathcal{M}$ for setting (n, Σ) with prior Q to be the expectation of the sum of payments to each agent and we write it as $AW_{\mathcal{M}}(n, \Sigma, Q, \mathbf{s})$. Note that for symmetric strategy profile, the **agent welfare** is proportional to each agent's expected payment since everyone plays the same strategy. A **Bayesian Nash equilibrium** consists of a strategy profile $\boldsymbol{s} = (s_1, \ldots, s_n)$ such that no player wishes to change her strategy, given the strategies of the other players and the information contained in the prior and her signal. See formal definitions in the full version of this paper.

2.3 Mechanism Design Tools: *f*-divergence and Proper Scoring Rules

Now we introduce f-divergence and strictly proper scoring rules, which are two of the main tools we will use in our mechanism design. f-divergence ([2, 5]) is always used in measuring the "difference" between distributions. One important property of the f-divergence family is information monotonicity: for any two distributions, if we use the same way to post-process each distribution, the two distributions will become "closer" because of potential information loses.

f-divergence

f-divergence [2, 5] $D_f : \Delta_\Sigma \times \Delta_\Sigma \to \mathbb{R}$ is a non-symmetric measure of difference between distribution $\mathbf{p} \in \Delta_\Sigma$ and distribution $\mathbf{q} \in \Delta_\Sigma$ and is defined to be

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{\sigma \in \Sigma} \mathbf{p}(\sigma) f\left(\frac{\mathbf{p}(\sigma)}{\mathbf{q}(\sigma)}\right)$$

where $f(\cdot)$ is a convex function.

We introduce three properties of f-divergence:

(1) Information Monotonicity [5, 3]: For any $\mathbf{p}, \mathbf{q}$, and transition matrix $\theta \in \mathbb{R}^{|\Sigma| \times |\Sigma|}$ where $\theta(\sigma, \sigma')$ is the probability that we map σ' to σ, we have $D(\mathbf{p}, \mathbf{q}) \geq D^*(\theta\mathbf{p}, \theta\mathbf{q})$. When θ is a permutation matrix θ_π, $D(\mathbf{p}, \mathbf{q}) = D(\theta_\pi \mathbf{p}, \theta_\pi \mathbf{q})$. When θ is not a permutation, the inequality is strict if $\mathbf{p}$ and $\mathbf{q}$ satisfy some weak conditions. The weak conditions are closely related to the definition of fine-grained prior (see Appendix C for more details.).

(2) Non-negative [5] For any $\mathbf{p}, \mathbf{q}$, $D_f(\mathbf{p}, \mathbf{q}) \geq 0$ and $D_f(\mathbf{p}, \mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$.

(3) Convexity [5]: Both $D_f(\cdot, \mathbf{q})$ and $D_f(\mathbf{p}, \cdot)$ are convex functions for any $\mathbf{p}, \mathbf{q}$.

Hellinger-divergence

If we pick the convex function $f(\cdot)$ as $(\sqrt{x} - 1)^2$, we will obtain Hellinger-divergence [5]

$$D^*(\mathbf{p}, \mathbf{q}) = \sum_{\sigma} (\sqrt{\mathbf{p}(\sigma)} - \sqrt{\mathbf{q}(\sigma)})^2$$

Thus Hellinger-divergence is a type of f-divergence.

In addition to the above three properties, we highlight two important properties Hellinger-divergence has:

(4) Square root triangle inequality [5]: $|\sqrt{D^*(\mathbf{p}, \mathbf{q})} - \sqrt{D^*(\mathbf{p}, \mathbf{q}')}| < \sqrt{D^*(\mathbf{q}', \mathbf{q})}$ for any $\mathbf{p}, \mathbf{q}, \mathbf{q}'$

(5) Bounded divergence [5]: $0 \leq D^*(\mathbf{p}, \mathbf{q}) \leq 1$

Proper Scoring Rules

Now we introduce strictly proper scoring rules, another key tool we will use in our mechanism design. Starting with [19], proper scoring rules have become a common ingredient in mechanisms for unverifiable information elicitation (e.g. [20, 27]).

A scoring rule $PS : \Sigma \times \Delta_\Sigma \to \mathbb{R}$ takes in a signal $\sigma \in \Sigma$ and a distribution over signals $\delta_\Sigma \in \Delta_\Sigma$ and outputs a real number. A scoring rule is *proper* if, whenever the first input is drawn from a distribution δ_Σ, then the expectation of PS is maximized by δ_Σ. A scoring rule is called *strictly proper* if this maximum is unique. We will assume throughout that the scoring rules we use are strictly proper. By slightly abusing notation, we can extend a scoring rule to be $PS : \Delta_\Sigma \times \Delta_\Sigma \to \mathbb{R}$ by simply taking $PS(\delta_\Sigma, \delta'_\Sigma) = \mathbb{E}_{\sigma \leftarrow \delta_\Sigma}(\sigma, \delta'_\Sigma)$. We note that this means that any proper scoring rule is linear in the first term.

▶ **Example 1** (Example of Proper Scoring Rule)**.** Fix an outcome space Σ for a signal σ. Let $\mathbf{q} \in \Delta_\Sigma$ be a reported distribution. The Logarithmic Scoring Rule maps a signal and reported distribution to a payoff as follows:

$$L(\sigma, \mathbf{q}) = \log(\mathbf{q}(\sigma)).$$

Let the signal σ be drawn from some random process with distribution $\mathbf{p} \in \Delta_\Sigma$.
Then the expected payoff of the Logarithmic Scoring Rule

$$\mathbb{E}_{\sigma \leftarrow \mathbf{p}}[L(\sigma, \mathbf{q})] = \sum_\sigma \mathbf{q}(\sigma) \log \mathbf{q}(\sigma) = L(\mathbf{p}, \mathbf{q})$$

According to [10], this value will be maximized if and only if $\mathbf{q} = \mathbf{p}$.

Proper scoring rules are a key tool in the design of mechanisms [20] in the BTS framework. In such mechanism, agents are asked to report their private information and forecast for other agents and paid based on a “prediction score” and an “information score”. The prediction score is usually calculated by a proper scoring rule and the information score is customized.

Prediction Score via Proper Scoring Rules. Agents will receive a prediction score based on how well their prediction predicts a randomly chosen agent's reported signal. Say an agent i reports prediction $\hat{\mathbf{p}}_i$ then a random agent, call him agent j, is chosen, agent i will receive a prediction score $PS(\hat{\sigma}_j, \hat{\mathbf{p}}_i)$ where PS is a proper scoring rule. Note that any proper scoring rule works. $PS(\hat{\sigma}_j, \hat{\mathbf{p}}_i)$ is maximized if and only if agent i's reported prediction $\hat{\mathbf{p}}_i$ is his expected likelihood for $\hat{\sigma}_j$. Agent i cannot pretend to have a different expected likelihood without reducing his expectation for his prediction score.

3 The Disagreement Mechanism

3.1 Buiding Block – Divergence-Based BTS

In this section, we introduce a building block of our Disagreement Mechanism – Divergence-Based BTS [22]. It follows the BTS framework and still pays agents an “information score” and a “prediction score”. The main idea of Divergence-Based BTS is that the mechanism punishes the **Inconsistency** of agents – the “difference” between two random agents' predictions when they report the same signal. The common prior assumption tells us agents cannot agree to disagree. That is, if two agents receive the same private information, they must have the same “belief” about the world. In our setting, if agents tell the truth, whenever two agents report the same signal, they will report the same prediction as well. Thus, everyone telling the truth is a consistent strategy. Since Divergence-Based BTS punishes inconsistency, the truth-telling strategy will be encouraged in Divergence-Based BTS.

Divergence-Based BTS [22] $\mathcal{M}$:

Let $\alpha, \beta > 0$ be parameters and let PS be a strictly proper scoring rule, then we define $\mathcal{M}(\alpha, \beta, PS)^2$ as follows:

1. Each agent i reports a signal and a prediction $r_i = (\hat{\sigma}_i, \hat{\mathbf{p}}_i)$
2. For each agent i and agent j, we define a prediction score that depends on agent i's prediction and agent j's report signal

$$score_P(r_i, r_j) = PS(\hat{\sigma}_j, \hat{\mathbf{p}}_i),$$

and an information score

$$score_I(r_i, r_j) = \begin{cases} 0 & \hat{\sigma}_i \neq \hat{\sigma}_j \\ -(PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_j) - PS(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_i)) & \hat{\sigma}_i = \hat{\sigma}_j \end{cases}$$

3. Each agent i is matched with a random agent j. The payment for agent i is

$$payment_{\mathcal{M}(\alpha,\beta,\mathcal{PS})}(i, \mathbf{r}) = \alpha score_P(r_i, r_j) + \beta score_I(r_i, r_j).$$

▶ **Theorem 2.** *[22] For any $\alpha, \beta > 0$ and any strictly proper scoring rule PS, $\mathcal{M}(\alpha, \beta, PS)$ has truth-telling as a strict Bayesian-Nash equilibrium whenever the prior Q is informative and symmetric.*

Main Drawback of Divergence-Based BTS

The main drawback is that there may be many other equilibria that have the same payoff with truth-telling. Agents can simply report the a priori most popular signal and predict that everyone does the same. This strategy is a consistent equilibrium and gives agents the maximum possible payoff since their predictions are perfect. In particular, for any non-trivial prior, this strategy pays *strictly more* than the truth-telling equilibrium – so that it Pareto dominates truth-telling.

The above extreme example provides a effortless and meaningless equilibrium but is preferred by agents in Divergence-Based BTS. To deal with this problem, one key observation is that in the meaningless equilibrium mentioned above, the unitary predictions implies their report profiles have little information. At a high level, the "disagreement" between agents represents the amount of information their report profiles have. Motivated by this extreme example, we design a new mechanism – the Disagreement Mechanism – that encourages "disagreement".

3.2 The Disagreement Mechanism and Main Theorem

In this section, we will describe our Disagreement Mechanism and state our main theorem. To design our mechanism, we start with the Divergence-Based BTS and (a) first use a typical trick to create a zero-sum game which has the same equilibria as the Divergence-Based BTS; (b) pay each agent an extra score that only depends on other agents which will not change the structure of the equilibria. We want this extra score to represent "classification score" (See Figure 1).

---

[2] This mechanism is a little bit different from Divergence-Based BTS mechanism [22]. Divergence-Based BTS uses specific proper scoring rule (log scoring rule). But it is easy to see using general proper scoring rules still keeps the strictly truthful property of Divergence-Based BTS. We defer the proof to the full version.

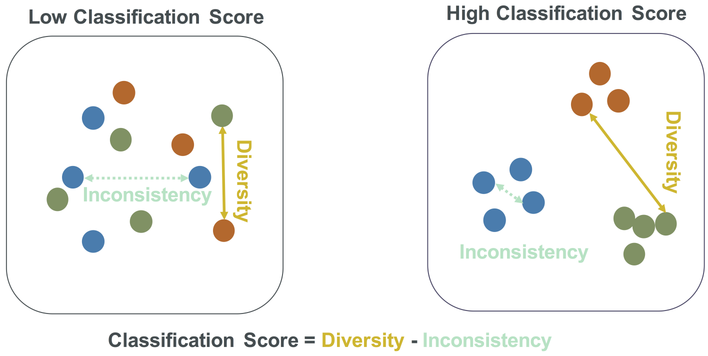

Figure 1 Illustration for Classification Score: Each point represents an agent's report profile – the *color* represents the *signal* the agent reports; the *position* represents the *prediction* the agent reports. We informally define **Inconsistency** as the *average* disagreement between every two agents' predictions when they report the *same* signal and **Diversity** as the *average* disagreement between every two agents' predictions when they report *different* signals. We informally define **Classification Score** as Diversity minus Inconsistency. Note that the report profiles in the right figure will have a much higher classification score than those in the left figure since the right figure has high Diversity and low Inconsistency.

Disagreement Mechanism $\mathcal{M}^+(\alpha, \beta, PS(\cdot,\cdot))$

$\mathbf{r} = \{r_1, r_2, ..., r_n\}$ is all agents' report profiles where for any r, $r_i = (\hat{\sigma}_i, \hat{\mathbf{p}}_i)$.

1. *Zero-sum Trick*: Divide the agents arbitrarily into two groups – group A and group B – such that both A and B have at least 3 agents. Each group of agents plays the game (mechanism) $\mathcal{M}$ that is restricted in their own group. For group A, each agent i_A receives a

$$\begin{aligned} score_{\mathcal{M}}(i_A, \mathbf{r}) &= payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i_A, \mathbf{r}_A) \\ &\quad - \frac{1}{|A|} \sum_{j_B \in B} payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(j_B, \mathbf{r}_B) \end{aligned}$$

Where $payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(i_A, \mathbf{r}_A)$ is agent i_A's payment when he is paid by mechanism $\mathcal{M}(\alpha, \beta, PS(\cdot,\cdot))$ given group A's report profiles $\mathbf{r}_A$ and that he can only be paired with a random peer from group A (we have similar explanation for $payment_{\mathcal{M}(\alpha,\beta,PS(\cdot,\cdot))}(j_B, \mathbf{r}_B)$). For agents in group B, we use the analogous way to score them.

2. *Additional Classification Reward*: Each agent i is matched with two random agents $j, k \neq i$ chosen from all agents (including group A and group B), the payment for agent i is

$$payment_{\mathcal{M}^+(\alpha,\beta,PS(\cdot,\cdot)}(i, \mathbf{r}) = score_{\mathcal{M}}(i, \mathbf{r}) + score_C(r_j, r_k)$$

where

$$score_C(r_j, r_k) = \begin{cases} D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) & \hat{\sigma}_j \neq \hat{\sigma}_k \\ -\sqrt{D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)} & \hat{\sigma}_j = \hat{\sigma}_k \end{cases}$$

recall that D^* denotes the Hellinger Divergence.

▶ **Theorem 3** (Main Theorem). *For any number of signals m, given any SNIFE prior, if the number of agents $n \geq 6$, then in $\mathcal{M}^+(\alpha, \beta, PS(\cdot,\cdot))$ with $\frac{\alpha}{\beta} < \frac{1}{4m}$,*

1. *(*Strictly truthful*) truth-telling is a strict Bayesian Nash equilibrium;*
2. *(*Focal*) in any permutation equilibrium,* every agent *has equal expected payment with truth-telling; and in any* symmetric *equilibrium that is not a permutation equilibrium,* every agent*'s expected payment is strictly less than that of truth-telling.*
3. *(*Robust Focal*) any* symmetric *equilibrium that pays within γ_1 of truth-telling* [3] *must be $\tau_1(\gamma_1)$ close to a permutation strategy profile; and moreover*
4. *(*Tight*) no detail free mechanism can have truth-telling as an equilibrium that has strictly higher agent-welfare than all other permutation equilibria.*

where $\tau_1(\gamma_1) = O(\sqrt[3]{\gamma_1})$, (the constants we omit only depend on the first two moments of prior Q)[4]*.*

We extend our results to asymmetric equilibria when the number of agents is sufficiently large in the full version.

3.3 Proof Highlights

In this section we give a few proof highlights.

First note that to show each agent's expected payment in a symmetric equilibrium is less than that of truth-telling, we only need to show the sum of all agents' expected payments – agent welfare – is less than that of truth-telling since everyone plays the same strategy in a symmetric equilibrium.

We first show that the agent welfare of our *Disagreement Mechanism* is the *Classification Score*, which follows by a straightforward computation. It remains to show that *Classification Score* has the aforementioned properties.

Best Prediction Strategy Profiles: We call a strategy profile a *best prediction strategy profile* if for any i, agent i reports a prediction that maximizes his prediction score. By some calculations, we know agent i's *best prediction* is $\theta_{-i}\mathbf{q}_{\sigma_i}$ given σ_i is his private signal and recall that $\theta_{-i} = \frac{\sum_{j\neq i}\theta_i}{n-1}$ where $(\theta_1, \theta_2, ..., \theta_n)$ is the signal strategy. We call this strategy profile a *symmetric* best prediction strategy profile if there exists a signal strategy θ such that $\theta_i = \theta$ for any i. Based on the definition of permutation strategy profile, it is clear that any permutation strategy profile is a symmetric *best prediction* strategy profile.

Consider two agents who report different signals. If they use a permutation strategy profile π whose signal strategy is θ_π, then their predictions will be $\theta_\pi\mathbf{q}_\sigma, \theta_\pi\mathbf{q}_{\sigma'}$ given their private signals are $\sigma \neq \sigma'$. If they use a symmetric best prediction strategy, then their reported predictions will be $\theta\mathbf{q}_\sigma, \theta\mathbf{q}_{\sigma'}$. In the first case, the Hellinger divergence between the two agents' reported predictions is $D^*(\theta_\pi\mathbf{q}_\sigma, \theta_\pi\mathbf{q}_{\sigma'}) = D^*(\mathbf{q}_\sigma, \mathbf{q}_{\sigma'})$ while in the second case, the Hellinger divergence between the two agents' reported predictions is $D^*(\theta\mathbf{q}_\sigma, \theta\mathbf{q}_{\sigma'}) \leq D^*(\mathbf{q}_\sigma, \mathbf{q}_{\sigma'}) = D^*(\theta_\pi\mathbf{q}_\sigma, \theta_\pi\mathbf{q}_{\sigma'})$. The inequality follows from the information monotonicity of Hellinger divergence. Thus, the two agents' predictions in the second case is "closer" than those in the first case. So a permutation strategy profile is more diverse than any other symmetric best prediction strategy, and additionally has no inconsistency. To make permutation strategy profiles beat symmetric best prediction strategy profiles, it is enough to just pay agents the additional diversity reward.

[3] Note that it is an additive gap.
[4] Actually $\tau_1(\gamma_1) = \frac{1}{c_1}\sqrt[3]{\frac{\gamma_1}{c_2c_3c_4}}$

General Equilibria: However, **the biggest challenge** is that there exists equilibria that are not best prediction strategy profiles. Thus, *it is not enough to just pay agents an additional diversity reward.* To deal with this challenge, we replace diversity by *classification score.* To show that classification score works, we map each equilibrium $\boldsymbol{s}^*$ to a strategy profile $\boldsymbol{s}^*_{BP}$ that belongs to *best prediction strategy profiles.* The *technical heart* of the proof bounds the classification score of an equilibrium strategy profile $\boldsymbol{s}^*$ by the diversity of its corresponding best prediction strategy profile $\boldsymbol{s}^*_{BP}$. Once we finish this, we can bound the classification score of any equilibrium strategy profile by the classification score of permutation strategy profiles (note that for permutation strategy profiles, the classification score is equal to the diversity since they are consistent strategy profiles) and complete the proof.

Arriving at this bound requires a non-trivial understanding of the structure of the equilibria, and especially the relation between the different agents' prediction reports in any equilibria. Given a strategy profile for the reports, we obtain a system of linear equations relating the prediction reports. Achieving this bound also requires the delicate use of the triangle inequality applied to $\sqrt{D^*}$. That is why we pick the Hellinger divergence rather than any other f-divergence.

Basically, considering that agents lie for the signal reports, we will show that it's better for them to report their best predictions. Moreover, the information monotonicity shows that even when the agents report their best predictions, it is still worse than truth-telling. Thus, our mechanism is focal.

Asymmetric Equilibria: In the more complicated asymmetric case, the difficulty is that even if agents play best prediction strategy profiles, we cannot use information monotonicity to prove permutation strategy profiles gain the strictly highest classification score. However, if the number of agents is large enough, we will see any strategy profile that belongs to *best prediction strategy profiles* family is "almost symmetric". Using "almost symmetric" result, we can generalize the above framework to approximate work for asymmetric case.

Finally, we show that equilibrium that having the classification score close to that of truth-telling, must be close to a permutation equilibrium.

Tightness Result: The intuitive explanation for this tightness result is that the agents can collude to relabel the signals and the mechanism has no way to defend against this relabelling without knowing some information about agents' common prior. The key idea to prove that result is what we refer to as **Indistinguishable Scenarios**, that is, for the scenario A where agents collude to relabel the signals, there always exists another scenario B where agents tell the truth such that no detail free and truthful mechanism can distinguish A and B.

3.4 Proof Outline for Main Theorem

In this section, we give the proof outline for our main theorem. Recall that we informally defined *Inconsistency* as the average disagreement between every two agents' predictions when they report the same signal and *Diversity* as the average disagreement between every two agents' predictions when they report different signals. We also defined *Classification Score* as Diversity minus Inconsistency. Here we give technical definitions of those concepts.

We first introduce a short hand which will simplify the formula for *Diversity* and *Inconsistency.*

$$\int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) \triangleq \int_{\hat{\sigma}_j,\hat{\mathbf{p}}_j,\hat{\sigma}_k,\hat{\mathbf{p}}_k} Pr_{(\hat{\sigma}_j,\hat{\mathbf{p}}_j)\leftarrow s_j(\sigma_j)}(\hat{\sigma}_j,\hat{\mathbf{p}}_j) Pr_{(\hat{\sigma}_k,\hat{\mathbf{p}}_k)\leftarrow s_k(\sigma_k)}(\hat{\sigma}_k,\hat{\mathbf{p}}_k)$$

where s_j is the strategy of agent j and $s_j(\sigma_j)$ is a distribution over agent j's report profile $(\hat{\sigma}_j, \hat{\mathbf{p}}_j)$ given agent j receives private signal σ_j and uses strategy s_j, and similarly for agent k. This defines the natural measure on the reports of agents j and k given that they play strategies s_j and s_k and a fixed prior Q (which is implicit), and allows us to succinctly describe probabilities of events in this space.

We define *Diversity* as the expected Hellinger divergence D^* between two random agents when they report different signals, so

$$\mathit{Diversity} = \sum_{\substack{j\\k\neq j}} \sum_{\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k})\delta(\hat{\sigma}_j \neq \hat{\sigma}_k) D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)$$

where $Pr(j,k)$ is the probability agents j,k are picked, and $Pr(\sigma_j,\sigma_k)$ is the probability that agent j receives private signal σ_j and agent k receives private signal σ_k.

Similarly, we can write down the technical definition for *Inconsistency*. But here we do not use Hellinger divergence as the "difference" function in $\sum_{u,v\in U, \mathcal{C}_r(u)=\mathcal{C}_r(v)} D(u,v)$, we use square root of the Hellinger divergence which is the Hellinger distance as the "difference" function. The reason is we want to use the convexity of the Hellinger divergence and the triangle inequality of the Hellinger distance. We will describe the details in the future. For now we give a technical definition for *Inconsistency*:

$$\mathit{Inconsistency} = -\sum_{\substack{j\\k\neq j}} \sum_{\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k})\delta(\hat{\sigma}_j = \hat{\sigma}_k) \sqrt{D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)}$$

In addition to the *Diversity* and *Inconsistency*, we also introduce a new concept – *TotalDivergence* – and then we use this value as a bridge. Recall that we defined $\mathit{Diversity} = \sum_{j,k\neq j,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k})\delta(\hat{\sigma}_j \neq \hat{\sigma}_k) D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)$, now we define a similar concept

$$\mathit{TotalDivergence} = \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k) \int_{\hat{j},\hat{k}} Pr(\hat{j},\hat{k}) D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k)$$

First note that total divergence is robust to summing over j,k or $j \neq k$ since when $j = k$, $D^*(\hat{\mathbf{p}}_j, \hat{\mathbf{p}}_k) = 0$.

We can see $\mathit{TotalDivergence} \geq \mathit{Diversity}$ since *TotalDivergence* also includes the divergence between the agents who report the same signals. We show that the equality holds if and only if $\mathit{Inconsistency} = 0$:

▶ **Claim 4.** *For any strategy profile* $\mathbf{s}$*,* $\mathit{Diversity}(\mathbf{s}) = \mathit{TotalDivergence}(\mathbf{s}) \Leftrightarrow \mathit{Inconsistency}(\mathbf{s}) = 0$

The above claim implies the below claim. We defer the proofs to the full version.

▶ **Claim 5.**

$$\begin{aligned}\mathit{ClassificationScore(truthtelling)} &= \mathit{Diversity(truthtelling)}\\ &= \mathit{TotalDivergence(truthtelling)}\end{aligned}$$

Now we begin to state our proof outline: For any equilibrium $\boldsymbol{s}$, we define a modified strategy for $\boldsymbol{s}$:

We define $\boldsymbol{s}_{BP}$ what we call a *best prediction strategy* of $\boldsymbol{s}$ as a strategy where each agent uses the same signal strategy which he uses in $\boldsymbol{s}$ but plays his *best prediction* which maximizes the prediction score. In this case, by some calculations (see full version for a detailed proof), for any i, agent i plays $\theta_{-i}\mathbf{q}_{\sigma_i}$. In the symmetric case, agent i play $\theta\mathbf{q}_{\sigma_i}$.

The results or our main theorem follows from two technical lemmas:

(1) ***ClassificationScore***$(s) \leq$ ***TotalDivergence***(s_{BP}). [Main Lemma: Lemma 7] This is our main lemma and the main technical ingredient we use to show our main lemma is the triangle inequality of the square root of Hellinger divergence. Once we show it, we can directly prove the focal property of the Disagreement Mechanism: considering that agents lie for the signal reports, our main technical lemma shows that it's better for them to report their best predictions. Moreover, the information monotonicity shows that even when the agents report their best predictions, it is still worse than truth-telling.
Note that this result is valid for any equilibrium $\boldsymbol{s}$ – symmetric or asymmetric – and is still a main ingredient when we extend the focal property to the asymmetric case.

(2) ***TotalDivergence***$(truthtelling) \approx$ ***TotalDivergence***$(s_{BP}) \Rightarrow \theta \approx \pi$ when $\boldsymbol{s}$ is a symmetric equilibrium with signal strategy θ. This, informally, means that if a symmetric equilibrium pays close to truth-telling, it must be close to a permutation equilibrium, and thus pays about the same as truth-telling.

3.5 Proof for Main Theorem

In this section, we are going to show the first two parts of Theorem 3. We defer the proof for other parts and the asymmetric extension to the full version.

Proof of Theorem 3 Part 1: $\mathcal{M}^+(\alpha, \beta, PS(\cdot,\cdot))$ is strictly truthful.

▶ **Lemma 6.** *The* Disagreement Mechanism *has the same equilibria as the Divergence-based BTS.*

We defer the proof of this lemma to the full version. Radanovic and Faltings [22] have already show $\mathcal{M}(\alpha, \beta, PS(\cdot,\cdot))$ has truth-telling as a strict equilibrium for any SNIFE prior in Theorem 2. Since $\mathcal{M}^+(\alpha, \beta, PS(\cdot,\cdot))$ does not change the equilibrium structure of $\mathcal{M}(\alpha, \beta, PS(\cdot,\cdot))$ according to Claim 6, we have $\mathcal{M}^+(\alpha, \beta, PS(\cdot,\cdot))$ has truth-telling as a strict equilibrium for any SNIFE prior as well. ◀

Proof of Theorem 3 Part 2: $\mathcal{M}^+(\alpha, \beta, PS(\cdot,\cdot))$ has truth-telling as a focal equilibrium. We use our main lemma *ClassificationScore*$(\boldsymbol{s}) <$ *TotalDivergence*$(\boldsymbol{s}_{BP})$ directly to prove: **any *symmetric* non-permutation equilibrium's agent welfare (*ClassificationScore*) must be strictly less than truth-telling**

▶ **Lemma 7** (Main Lemma)**.** *For any equilibrium $\boldsymbol{s}$, if $\boldsymbol{s}_{BP}$ is a best prediction strategy of $\boldsymbol{s}$, we have*

$$\textit{ClassificationScore}(\boldsymbol{s}) \leq \textit{TotalDivergence}(\boldsymbol{s}_{BP})$$

If the equality holds, then we have Inconsistency$(\boldsymbol{s}) = 0$ and $s = \boldsymbol{s}_{BP}$.

We defer the proof of our main lemma to the full version. Notice that if all agents play a symmetric signal strategy θ, then for any j, k, $\theta_{-j} = \theta_{-k} = \theta$. For any symmetric non-permutation equilibrium $\boldsymbol{s}$, it is possible that the signal strategy of $\boldsymbol{s}$ is not a permutation or it is a permutation θ_π but agents do not report $\theta_\pi\mathbf{q}_\sigma$ given σ is their private signal. So we consider two cases:

(a) We first consider the case that the signal strategy θ of $\boldsymbol{s}$ is a permutation matrix θ_π, but agents do not report $\theta_\pi \mathbf{q}_\sigma$. That is, agents collude to relabel the signal but do not report the best predictions based on the collusion.

$$\begin{aligned}
\mathit{ClassificationScore}(\boldsymbol{s}) &< \mathit{TotalDivergence}(\boldsymbol{s}_{BP}) \\
&= \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta_{-j}\mathbf{q}_{\sigma_j}, \theta_{-k}\mathbf{q}_{\sigma_k}) \\
&\qquad \text{(Definition of best prediction strategy)} \\
&= \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta_\pi\mathbf{q}_{\sigma_j}, \theta_\pi\mathbf{q}_{\sigma_k}) \\
&\qquad \text{(In Case (a), the signal strategy } \theta \text{ of } \boldsymbol{s} \text{ is a permutation matrix } \theta_\pi\text{)} \\
&= \sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\mathbf{q}_{\sigma_j}, \mathbf{q}_{\sigma_k}) \\
&= \mathit{TotalDivergence}(\mathit{truthtelling}) \\
&= \mathit{ClassificationScore}(\mathit{truthtelling})
\end{aligned}$$

The first inequality follows from our main lemma. The inequality is strict for the following reason: when the signal strategy θ of $\boldsymbol{s}$ is a permutation matrix, $\boldsymbol{s}_{BP}$ is a permutation strategy profile. Based on our main lemma if $\mathit{ClassificationScore}(\boldsymbol{s}) = \mathit{TotalDivergence}(\boldsymbol{s}_{BP})$, we have $\boldsymbol{s} = \boldsymbol{s}_{BP}$ which implies that $\boldsymbol{s}$ is a permutation strategy profile which is a contradiction to the fact $\boldsymbol{s}$ is a non-permutation strategy profile.
The last equality follows from Corollary 5.

(b) We consider the case that the signal strategy θ of $\boldsymbol{s}$ is not a permutation matrix. In this case, agents may or may not report their best predictions. Our main technical lemma shows that it's better for them to report their best predictions. The information monotonicity shows that even when the agents report the best predictions, it is still worse than truth-telling. We still use case (a)'s proof. Even though the inequality in the first line may not be strict, the equality in the fourth line must be a strict inequality:

$$\begin{aligned}
&\sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\theta\mathbf{q}_{\sigma_j}, \theta\mathbf{q}_{\sigma_k}) \\
< &\sum_{j,k,\sigma_j,\sigma_k} Pr(j,k)Pr(\sigma_j,\sigma_k)D^*(\mathbf{q}_{\sigma_j}, \mathbf{q}_{\sigma_k}) \qquad \text{(Information Monotonicity)}
\end{aligned}$$

since based on Corollary 16, when θ is not a permutation, and Q is fine-grained and non-zero, the information monotonicity is strict. So in both of the above two cases, we have

$$\mathit{ClassificationScore}(\boldsymbol{s}) < \mathit{ClassificationScore}(\mathit{truthtelling})$$

if $\boldsymbol{s}$ is not a permutation equilibrium. Therefore, our mechanism is focal. ◀

Acknowledgements. We thank David Parkes for helpful conversations and suggestions.

References

1 Arpit Agarwal, Debmalya Mandal, David C. Parkes, and Nisarg Shah. Peer prediction with heterogeneous users. In Constantinos Daskalakis, Moshe Babaioff, and Hervé Moulin, editors, *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017*, pages 81–98. ACM, 2017. doi:10.1145/3033274.3085127.

2 Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

3 S-I Amari and A Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

4 Yang Cai, Constantinos Daskalakis, and Christos H Papadimitriou. Optimum statistical estimation with strategic data sources. *arXiv preprint arXiv:1408.2539*, 2014.

5 Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

6 Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pages 319–330. International World Wide Web Conferences Steering Committee, 2013.

7 Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. Incentives to counter bias in human computation. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

8 Rafael M Frongillo and Jens Witkowski. A geometric method to construct minimal peer prediction mechanisms. In *AAAI*, pages 502–508, 2016.

9 Xi Alice Gao, Andrew Mao, Yiling Chen, and Ryan Prescott Adams. Trick or treat: putting peer prediction to the test. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 507–524. ACM, 2014.

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

11 Radu Jurca and Boi Faltings. Collusion-resistant, incentive-compatible feedback payments. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 200–209. ACM, 2007.

12 Radu Jurca and Boi Faltings. Mechanisms for making crowds truthful. *J. Artif. Int. Res.*, 34(1), 2009.

13 Vijay Kamble, Nihar Shah, David Marn, Abhay Parekh, and Kannan Ramachandran. Truth serums for massively crowdsourced evaluation tasks. *arXiv preprint arXiv:1507.07045*, 2015.

14 Y. Kong and G. Schoenebeck. A Framework For Designing Information Elicitation Mechanisms That Reward Truth-telling. *ArXiv e-prints*, 2016. arXiv:1605.01021.

15 Y. Kong and G. Schoenebeck. Equilibrium Selection in Information Elicitation without Verification via Information Monotonicity. *ArXiv e-prints*, mar 2016. arXiv:1603.07751.

16 Y. Kong, G. Schoenebeck, and K. Ligett. Putting Peer Prediction Under the Micro(economic)scope and Making Truth-telling Focal. *ArXiv e-prints*, mar 2016. arXiv:1603.07319.

17 Yang Liu and Yiling Chen. Machine-learning aided peer prediction. In Constantinos Daskalakis, Moshe Babaioff, and Hervé Moulin, editors, *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017*, pages 63–80. ACM, 2017. doi:10.1145/3033274.3085126.

18 Debmalya Mandal, Matthew Leifer, David C Parkes, Galen Pickard, and Victor Shnayder. Peer prediction with heterogeneous tasks. *arXiv preprint arXiv:1612.00928*, 2016.

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

20 D. Prelec. A Bayesian Truth Serum for subjective data. *Science*, 306(5695):462–466, 2004.

21 Goran Radanovic and Boi Faltings. A robust bayesian truth serum for non-binary signals. In Marie desJardins and Michael L. Littman, editors, *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA.* AAAI Press, 2013. URL: `http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6451`.

22 Goran Radanovic and Boi Faltings. Incentives for truthful information elicitation of continuous signals. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

23 Goran Radanovic and Boi Faltings. Incentive schemes for participatory sensing. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1081–1089. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

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

25 V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes. Informed Truthfulness in Multi-Task Peer Prediction. *ArXiv e-prints*, mar 2016. `arXiv:1603.03151`.

26 Jens Witkowski, Bernhard Nebel, and David C Parkes. *Robust Peer Prediction Mechanisms.* PhD thesis, Ph. D. Dissertation, Albert-Ludwigs-Universitat Freiburg: Institut fur Informatik, 2014.

27 Jens Witkowski and David C. Parkes. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*, 2011.

28 Jens Witkowski and David C Parkes. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 964–981. ACM, 2012.

29 Jens Witkowski and David C Parkes. 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*, page 14. Citeseer, 2013.

30 Peter Zhang and Yiling Chen. Elicitability and knowledge-free elicitation with peer prediction. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 245–252. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

A Notation Table

B Formal Definitions of the SNIFE Priors

▶ **Assumption 8** (Symmetric Prior). *We assume throughout that the agents' signals* $\boldsymbol{\sigma}$ *are drawn from some joint* **symmetric prior** Q.

Because we will assume that the prior is symmetric, we denote $q_i(\sigma)$ by $q(\sigma)$ and $q_{i,j}(\sigma|\sigma')$ (where $i \neq j$) by $q(\sigma|\sigma')$. We also define $\mathbf{q}_\sigma = q(\cdot|\sigma)$.

▶ **Assumption 9** (Non-zero Prior). *We assume that for any* $\sigma, \sigma' \in \Sigma$, $q(\sigma) > 0, q(\sigma|\sigma') > 0$.

▶ **Assumption 10** (Informative Prior). *We assume if agents have different private signals, they will have different expectations for the fraction of at least one signal. That is for any* $\sigma \neq \sigma'$, *there exists* σ'' *such that* $q(\sigma''|\sigma) \neq q(\sigma''|\sigma')$.

The following assumption conceptually states that one state is not just a more likely version of another state, and can be thought of as a weaker version of assuming $q(\sigma|\cdot)$ are linearly independent.

Table 2 Basic Notations.

Σ	$\triangleq$	the space of all possible signals
Δ_Σ	$\triangleq$	the space of all probability distributions over Σ
σ_i	$\triangleq$	agent i's private signal
$\hat{\sigma}_i$	$\triangleq$	agent i's reported signal
$\mathbf{p}_i$ or $\mathbf{q}_{\sigma_i}$	$\triangleq$	agent i's prediction for other agents' private signals conditioning on she receives σ_i
$\hat{\mathbf{p}}_i$	$\triangleq$	agent i's reported prediction
$r_i = (\hat{\sigma}_i, \hat{\mathbf{p}}_i)$	$\triangleq$	agent i's report
$\mathbf{r}$	$\triangleq$	all agents' reports
s_i	$\triangleq$	agent i's strategy
θ_i	$\triangleq$	the signal strategy (marginal distribution) of s_i
θ_{-i}	$\triangleq$	$\frac{\sum_{j\neq i}\theta_j}{n-1}$ the average signal strategy excluding θ_i
$\bar{\theta}_n$	$\triangleq$	$\frac{\sum_i \theta_i}{n}$ the average signal strategy
$\mathbf{s}$	$\triangleq$	the strategy profile agents have
Q	$\triangleq$	the common symmetric prior agents share
$q(\sigma'\|\sigma)$	$\triangleq$	an agent's expected fraction of other agents who receives σ' when this agent receives σ
$q(\sigma)$	$\triangleq$	the priori expected fraction of agents who will receive σ

► **Assumption 11** (Fine-grained Prior). *We assume that for any $\sigma \neq \sigma' \in \Sigma$, there exists σ'', σ''' such that*

$$\frac{q(\sigma|\sigma'')}{q(\sigma'|\sigma'')} \neq \frac{q(\sigma|\sigma''')}{q(\sigma'|\sigma''')}$$

If this assumption does not hold, then in some sense σ and σ' are the same signal since the agents who receive σ have the same posterior with the agents who receive σ'. Therefore, We can replace σ and σ' with a new signal $\sigma_0 := \sigma$ or σ', and not lose any information.

We illustrate this in the below example:

► **Example 12.** $Q = \begin{pmatrix} q(s_1|s_1) & q(s_1|s_2) & q(s_1|s_3) \\ q(s_2|s_1) & q(s_2|s_2) & q(s_2|s_3) \\ q(s_3|s_1) & q(s_3|s_2) & q(s_3|s_3) \end{pmatrix} = \begin{pmatrix} 0.1 & 0.2 & 0.3 \\ 0.2 & 0.4 & 0.6 \\ 0.7 & 0.4 & 0.1 \end{pmatrix}$ is not a fine-grained prior since

$$\frac{q(s_1|s_1)}{q(s_2|s_1)} = \frac{q(s_1|s_2)}{q(s_2|s_2)} = \frac{q(s_1|s_3)}{q(s_2|s_3)}$$

Note that in this example, even we combine s_1 and s_2 to be a single signal s_0 which is defined as $s_0 := s_1$ or s_2, we do not lose any information: if an agent knows that the fraction of agents who report s_0 is x, we know his belief for the expectation of the fraction of s_1 must be $\frac{x}{3}$ no matter what private signal he receives.

We only require the fine-grained prior assumption to show that truth-telling is *strictly* "better" than any other symmetric equilibrium (excluding permutation equilibrium). In the above example where the prior is not fine-grained, if agents always report s_1 when they receive s_1 or s_2, this does not lose information (is not "worse") comparing with the case agents always tell the truth. So we cannot say truth-telling is strictly "better" than any other equilibrium when the prior is not fine-grained. However, this assumption is not necessary to show that truth-telling is a strict Bayesian equilibrium of our mechanism, nor to show that the agent welfare of truth-telling is at least as high as other symmetric equilibrium.

▶ **Assumption 13** (Ensemble Prior). *Although we talk of a single prior, in fact we have an ensemble $Q = \{Q_n\}_{n\in N, n\geq 3}$ of priors; one for each possible number of agents greater than 3. We assume that all Q_n are over the same signal set Σ have identical $q(\sigma)$ and $q(\sigma'|\sigma)$.*

When the number of agents n changes, the joint prior actually changes as well, but the first two moments of the prior are fixed. This allows us to make meaningful statements about n going to infinity.

C Information Monotonicity

▶ **Lemma 14** (Information Monotonicity ([3])). *For any strictly convex function f, f-divergence $D_f(\mathbf{p},\mathbf{q})$ satisfies information monotonicity so that for any transition matrix $\theta \in \mathbb{R}^{\Sigma\times\Sigma}$, $D_f(\mathbf{p},\mathbf{q}) \geq D_f(\theta\mathbf{p},\theta\mathbf{q})$.*

Moreover, the inequality is strict if and only if there exists σ,σ',σ'' such that $\theta(\sigma,\sigma')\mathbf{p}(\sigma') > 0$, $\theta(\sigma,\sigma'')\mathbf{p}(\sigma'') > 0$ and $\frac{\mathbf{p}(\sigma'')}{\mathbf{p}(\sigma')} \neq \frac{\mathbf{q}(\sigma'')}{\mathbf{q}(\sigma')}$.

We introduce the proof and give an example where the strictness condition is not satisfied here.

Proof. The proof follows from algebraic manipulation and one application of convexity.

$$
\begin{aligned}
D_f(\theta\mathbf{p},\theta\mathbf{q}) &= \sum_\sigma (\theta\mathbf{p})(\sigma) f\left(\frac{(\theta\mathbf{q})(\sigma)}{(\theta\mathbf{p})(\sigma)}\right) && (1)\\
&= \sum_\sigma \theta(\sigma,\cdot)\mathbf{p} f\left(\frac{\theta(\sigma,\cdot)\mathbf{q}}{\theta(\sigma,\cdot)\mathbf{p}}\right) && (2)\\
&= \sum_\sigma \theta(\sigma,\cdot)\mathbf{p} f\left(\frac{1}{\theta(\sigma,\cdot)\mathbf{p}}\sum_{\sigma'}\theta(\sigma,\sigma')\mathbf{p}(\sigma')\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')}\right) && (3)\\
&\leq \sum_\sigma \theta(\sigma,\cdot)\mathbf{p}\frac{1}{\theta(\sigma,\cdot)\mathbf{p}}\sum_{\sigma'}\theta(\sigma,\sigma')\mathbf{p}(\sigma') f\left(\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')}\right) && (4)\\
&= \sum_\sigma \mathbf{p}(\sigma) f\left(\frac{\mathbf{q}(\sigma)}{\mathbf{p}(\sigma)}\right) = D_f(\mathbf{p},\mathbf{q}) && (5)
\end{aligned}
$$

The second equality holds since $(\theta\mathbf{p})(\sigma)$ is dot product of the σ^{th} row of θ and $\mathbf{p}$.

The third equality holds since $\sum_{\sigma'}\theta(\sigma,\sigma')\mathbf{p}(\sigma')\frac{\mathbf{q}(\sigma')}{\mathbf{p}(\sigma')} = \theta(\sigma,\cdot)\mathbf{q}$.

The fourth inequality follows from the convexity of $f(\cdot)$.

The last equality holds since $\sum_\sigma \theta(\sigma,\sigma') = 1$.

We now examine under what conditions the inequality in Equation 4 is strict. Note that for any strictly convex function g, if $\forall u, \lambda_u > 0$, $g(\sum_u \lambda_u x_u) = \sum_u \lambda_u g(x_u)$ if and only if there exists x such that $\forall u, x_u = x$. By this property, the inequality is strict if and only if there exists σ,σ',σ'' such that $\frac{\mathbf{p}(\sigma'')}{\mathbf{p}(\sigma')} \neq \frac{\mathbf{q}(\sigma'')}{\mathbf{q}(\sigma')}$ and $\theta(\sigma,\sigma')\mathbf{p}(\sigma') > 0$, $\theta(\sigma,\sigma'')\mathbf{p}(\sigma'') > 0$. ◀

To understand the strictness condition more in Lemma 14, we give an example where the strictness condition is not satisfied:

▶ **Example 15.** $\mathbf{p} = (0.1\ 0.2\ 0.7)$, $\mathbf{q} = (0.2\ 0.4\ 0.4)$, $\theta = \begin{pmatrix} 0.3 & 0.6 & 0 \\ 0.7 & 0.4 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

We show by case analysis that we cannot find $\sigma, \sigma', \sigma''$ such that $\theta(\sigma, \sigma')\mathbf{p}(\sigma') > 0$, $\theta(\sigma, \sigma'')\mathbf{p}(\sigma'') > 0$ and $\frac{\mathbf{p}(\sigma'')}{\mathbf{p}(\sigma')} \neq \frac{\mathbf{q}(\sigma'')}{\mathbf{q}(\sigma')}$.

First note that because $\frac{\mathbf{p}(\sigma'')}{\mathbf{p}(\sigma')} \neq \frac{\mathbf{q}(\sigma'')}{\mathbf{q}(\sigma')}$, it cannot be that $\sigma' = \sigma''$, nor can it be the case that $\sigma', \sigma'' \in \{1, 2\}$ because $\frac{\mathbf{p}(1)}{\mathbf{p}(2)} = \frac{\mathbf{q}(1)}{\mathbf{q}(2)}$ and $\frac{\mathbf{p}(2)}{\mathbf{p}(1)} = \frac{\mathbf{q}(2)}{\mathbf{q}(1)}$. Thus it must be that either $\sigma' \in \{1, 2\}$ and $\sigma'' = 3$ or $\sigma' = 3$ and $\sigma'' \in \{1, 2\}$. Because these are symmetric, we consider the first case.

Because $\theta(\sigma, \sigma')\mathbf{p}(\sigma') > 0$ it must be that $\sigma \in \{1, 2\}$, but because $\theta(\sigma, \sigma'')\mathbf{p}(\sigma'') > 0$, it must be that $\sigma = 3$. So no assignment of $\sigma, \sigma', \sigma''$ is possible.

Thus, the strictness condition is not satisfied. By simple calculations, we have $\theta\mathbf{p} = (0.15\ 0.15\ 0.7)$, $\theta\mathbf{q} = (0.3\ 0.3\ 0.4)$. By some algebraic calculations, we have $D_f(\mathbf{p}, \mathbf{q}) = D_f(\theta\mathbf{p}, \theta\mathbf{q})$ for any function f.

▶ **Corollary 16.** *Given SNIFE prior Q, for any θ that is not a permutation, there exists two private signals $\sigma_1 \neq \sigma_2$ such that $D_f(\theta\mathbf{q}_{\sigma_1}, \theta\mathbf{q}_{\sigma_2}) < D_f(\mathbf{q}_{\sigma_1}, \mathbf{q}_{\sigma_2})$*

Proof. First notice that when θ is not a permutation, there exists a row of θ such that the row has at least two positive entries, in other words, there exists $\sigma, \sigma', \sigma''$ such that $\theta(\sigma, \sigma'), \theta(\sigma, \sigma'') > 0$. Based on the non-zero and fine-grained assumptions of Q, there exists $\sigma_1 \neq \sigma_2$ such that
$\theta(\sigma, \sigma')\mathbf{p}(\sigma'), \theta(\sigma, \sigma'')\mathbf{p}(\sigma'') > 0$ and $\frac{\mathbf{p}(\sigma')}{\mathbf{p}(\sigma'')} \neq \frac{\mathbf{q}(\sigma')}{\mathbf{q}(\sigma'')}$ where $\mathbf{p} = \mathbf{q}_{\sigma_1}, \mathbf{q} = \mathbf{q}_{\sigma_2}$. When $\theta(\sigma, \sigma')\mathbf{p}(\sigma'), \theta(\sigma, \sigma'')\mathbf{p}(\sigma'') > 0$, we have $\theta(\sigma, \cdot)\mathbf{p} > 0$. By Lemma 14, we have

$$D_f(\theta\mathbf{q}_{\sigma_1}, \theta\mathbf{q}_{\sigma_2}) < D_f(\mathbf{q}_{\sigma_1}, \mathbf{q}_{\sigma_2}).$$ ◀