Belief Elicitation from Agents with Competing Incentives

MANUEL WÜTHRICH, MARK YORK, DAVID PARKES

We study the problem of eliciting beliefs about an uncertain future event to inform a decision. Recommenders (belief-reporting agents) have a vested interest in the decision and may attempt to manipulate it by reporting untruthfully, a problem that is common in practice but has not been studied extensively. For instance, a climate scientist with ties to construction companies may overstate the risk of an extreme weather event to encourage investment into flood barriers. We build a model of this novel setting and study the extent to which recommenders manipulate the decision under different mechanisms (which include accuracy payments to recommenders). We quantify the lowest achievable manipulation under any mechanism depending on the (i) sensitivity of a target decision function to reports, (ii) magnitude of competing incentives, (iii) recommender compensation budget, and (iv) number of recommenders. We propose a mechanism that achieves these bounds (up to log factors) for a family of target decision functions. Finally, we show that if competing incentives are not intrinsic to recommenders, instead coming as bribes from a rational third party, then manipulation can be avoided entirely, with a sufficiently large mechanism budget, by making it too costly for the third party to bribe.

Authors affiliated with the Harvard University School of Engineering and Applied Sciences

1 INTRODUCTION

There are a wide range of settings where it is helpful, in making a decision, to elicit beliefs about the probability of an uncertain event. For example, suppose a city government must decide how much money to invest into flood defenses. To make an informed decision, the government may solicit predictions from experts (who we call *recommenders*), such as climate scientists, on whether an extreme weather event such as a hurricane will occur within the next five years. We assume the government, with accurate knowledge of these beliefs, has a well-defined method by which it desires to aggregate these beliefs and decide how much to invest into flood defenses. However, these beliefs are privately held, and the government must hence design a payment scheme to elicit accurate belief reports from recommenders and then make its investment decision based on these reports.

In the standard problem of *information elicitation with verification* [Brier et al., 1950], one elicits a belief from each participant and scores this report against the realization of a future event. This problem has been studied extensively and many payment schemes (*scoring rules*) that ensure truthful reporting of beliefs are known. However, these approaches do not address a fundamental problem: recommenders may have a vested interest in the decision taken on the basis of the reported beliefs. This interest may outweigh the incentive provided by the payment scheme and cause recommenders to seek to manipulate the decision by reporting untruthfully. In the above example, some climate scientists may own real estate that would lose value due to construction of nearby flood defense barriers, which may lead them to understate the risk. Others may be bribed by a construction company interested in building a flood barrier, which may lead them to overstate the risk.

In this paper, we study how such vested interests (which we call *competing incentives*) influence the reporting of recommenders and the decision made by the entity running the belief elicitation mechanism (the *principal*). We consider different methods for aggregating the true beliefs of recommenders into a decision (we call this the *target decision function*). For a given target decision function, we seek to design a *mechanism* that is able to realize (or *implement*) the target decision function, even though the decision function is defined on private beliefs while the mechanism must rely on reports produced by recommenders with competing incentives. A mechanism consists of a *decision rule*, which maps reports to a decision, and a *payment rule*, which maps reports and the realized outcome to a payment for each recommender. Our main results are:

(1) We show that for any nontrivial target decision function, no mechanism can exactly implement the target decision function when recommenders have intrinsic competing incentives. As a corollary, there are incentives to deviate from truthful reports in any mechanism that simply adopts the target decision function as its decision rule.
(2) We give a lower bound on the extent to which the actual decision taken, based on the reports, will deviate from the target decision function under *any mechanism*. For this purpose, we introduce a property of target decision functions, which we call *sensitivity*, and show that this property captures how susceptible any mechanism that seeks to approximate the decision function is to manipulation (i.e. deviation of the actual decision taken from the target decision function). This notion of sensitivity measures how responsive the target decision function is to changes in recommender beliefs.
(3) We introduce a novel mechanism that, given a target decision function, uses a payment rule that "concentrates incentives" where the target decision function is most sensitive to beliefs. We show that for any target decision function within some family of piecewise-linear functions, this mechanism achieves the lower bound on manipulation (up to a log factor).

(4) We also study the setting where the competing incentives are not intrinsic to recommenders, but instead stem from a bribe made by a *rational briber* who cares about influencing the decision in a particular direction and derives value from doing so (e.g., a construction company bribing a climate scientist). We prove that, if the principal's budget is large enough relative to the sensitivity of the target decision function and the value to the briber in changing the decision, our novel mechanism avoids manipulation entirely by making bribes too ineffective for the briber.

1.1 Overview of the Results

To illustrate our results, we return to the example of constructing a flood barrier.

Example 1. *A city government (principal) must decide how much money ($A \geq 0$) to invest into flood barriers. To make this decision, the principal wants to assess the probability of an extreme weather event within the next five years ($O = 1$) or not ($O = 0$) by consulting $n \geq 1$ experts (recommenders). Each recommender has a private estimate $Q_i \in [0,1]$ of the probability of $O = 1$, which the principal asks them to report (let $R_i \in [0,1]$ be i's report), in exchange for a payment that will be made after observing the outcome O (the total budget of the principal is $\beta > 0$). The target decision function $f : [0,1]^n \to \mathbb{R}_+$ specifies how much the principal would invest into flood barriers if they had access to all recommenders' private beliefs, i.e. for beliefs $Q = (Q_1, .., Q_n)$ the principal would invest $f(Q) > 0$. For instance, the principal may want to take the average across recommenders' beliefs, $\bar{Q} = \frac{1}{n}\sum_{i=1}^n Q_i$, and invest proportionally to that estimated probability,*

$$f(Q) = \ell \cdot \bar{Q}, \tag{1}$$

where $\ell \geq 0$ is the slope of the target decision function and $\bar{f}$ is an offset. The difficulty is that the principal does not know $Q = (Q_1, .., Q_n)$, and must rely on reports $R = (R_1, \ldots, R_n)$, which may not be truthful due to recommenders' competing incentives. For instance, suppose recommender i has ties to a construction company, and stands to earn a value $C_i \cdot A > 0$ that is linear in the investment A made into flood barriers, where $C_i > 0$ determines the extent to which i benefits from A. Such a recommender may have an incentive to increase the investment, by reporting $R_i > Q_i$. Another recommender j may own real estate, and may incur loss $C_j \cdot A < 0$, where $C_i < 0$, proportional to the amount invested into flood barriers, which could lead j to report $R_j < Q_j$. For reports $R = (R_1, \ldots, R_n)$, the principal can design a mechanism with a decision rule, $F(R) \geq 0$, and a payment rule, $\text{PAY}_i(R, O)$, that makes a payment to each recommender i after observing whether an extreme weather event happened ($O = 1$) or not ($O = 0$). Can the principal design a mechanism that accurately follows the decision $f(Q)$ prescribed by the target decision function f, despite these competing incentives?

A mechanism here consists of a payment rule $\text{PAY} = (\text{PAY}_i)_{i\in[n]}$ for each recommender and a decision rule $\text{F} : [0,1]^n \to \mathbb{R}_+$ that maps any report profile R to a decision $A \geq 0$. To implement target decision function f, it would be natural to consider $\text{F} := f$. However, we shall see that for nontrivial f, no payment scheme can ensure truthful reporting $R = Q$. Hence, we allow the mechanism to also use some $\text{F} \neq f$ that may deviate from f but reduce the amount by which recommenders misreport or the effect of misreports. Hence, for full generality, we define a mechanism by $\text{MECH} = (\text{F}, \text{PAY})$, and allow any decision rule F.[1]

[1]The *revelation principle* Hurwicz [1960] suggests that, w.l.o.g., we can focus on truthful mechanisms, i.e., PAY and F that ensure that recommenders always choose to report $R = Q$. In this case, it would be w.l.o.g. to frame this study as one of understanding which target decision functions can be combined with a payment rule to make them incentive compatible (including approximate, truthful implementations in the sense of algorithmic mechanism design). Our focus, however, is on *indirect* mechanisms, because it is unrealistic in this setting to elicit the full type of a recommender, which would include, for example, any private data they may have, their beliefs about what data others may have, how they think all this data relates

1.1.1 Intrinsic Competing Incentives. We show that for any nontrivial target decision function f, there is no mechanism MECH that can ensure $f(Q) = \mathrm{F}(R)$ in the presence of competing incentives $C_i > 0$, and give a lower bound on the extent to which the realized decision will deviate from the decision selected by the target decision function. The following is an informal version of our main negative result:

Theorem 2. *(Informal): For any target decision function f and any mechanism MECH $= (F, \text{PAY})$ with budget $\beta > 0$, there exist recommender beliefs $Q = (Q_1, \ldots, Q_n)$ and competing incentives $C = (C_1, \ldots, C_n)$ (bounded by $\sum_{i=1}^n |C_i| \leq \gamma$) such that recommenders' Bayes-Nash equilibrium reports R satisfy*

$$|F(R) - f(Q)| \gtrapprox \frac{\gamma}{\beta} \cdot \text{SENSITIVITY}_f,$$

where SENSITIVITY_f is a measure of the sensitivity of f.

We will formally define the measure of sensitivity in the technical section of this paper. Informally, it measures how fast target function, f, varies as a function of beliefs, and only equals 0 if f completely ignores beliefs. Hence, if the sensitivity of f is nonzero and there is a competing incentive $\gamma > 0$, then no finite budget can guarantee that there will be no manipulation. Further, we see that the manipulation scales linearly with the ratio between the size of competing incentive γ and the budget of the mechanism β. Applying Theorem 2 to the target decision function in Example 1, and ignoring some constants and log factors, we obtain, under the conditions of Theorem 2, that for any mechanism (F, PAY) there exist beliefs Q and competing incentives C such that equilibrium reports R have the effect

$$|\mathrm{F}(R) - f(Q)| \gtrapprox \frac{\gamma}{\beta} \cdot \frac{\ell^2}{n},$$

where $\ell \geq 0$ is the slope of the target decision function ($\ell = 0$ there is zero sensitivity), and the effect of the manipulation increases quadratically with ℓ. This lower bound decreases with the number of recommenders, and we shall see that this is not an artifact of the bound: we give a mechanism that achieves this bound (up to factors that are logarithmic in n) for a family of piecewise-linear functions. The novelty of this mechanism, which we call the *adaptive-payment mechanism*, is that the payment rule concentrates incentives where the target decision function is most sensitive to reports. For the following positive result, rationalizability Bernheim [1984], Pearce [1984] as a solution concept, because it includes all strategies a rational agent could possibly follow. We will define this notion formally in the technical section of this paper. Hence, using our novel mechanism, we obtain the following result:

Theorem 3. *(Informal): For any target decision function f in some family of piecewise-linear functions, the adaptive-payment mechanism with budget $\beta > 0$ achieves the lower bound from Theorem 2, up to log factors. In particular, for the target decision function in Example 1, we obtain, for any beliefs Q, and for any rationalizable report profile R, that*

$$|F(R) - f(Q)| \lessapprox \frac{\gamma}{\beta} \cdot \frac{\ell^2}{n},$$

for competing incentives bounded by $\sum_{i=1}^n |C_i| \leq \gamma$, target decision function with slope $\ell \geq 0$, and n recommenders.

---

to the outcome O, their competing incentives, and so forth. Hence, the revelation principle does not apply, and we need to consider decision rules $\mathrm{F} \neq f$, and study $|\mathrm{F}(R) - f(Q)|$, where we also allow for the reports to deviate from true beliefs.

We see from Theorem 3 that there are three ways of decreasing the effect of manipulation in the context of these piecewise-linear target decision functions: (i) make the target decision function f less sensitive, by decreasing slope ℓ, which is however often not desirable, as this makes it less responsive to recommenders' beliefs, (ii) increase the mechanism's budget β; or, (iii) increase the number of recommenders n. The third option is perhaps the most interesting in practice, as it allows to maintain the sensitivity of the target decision function without having to increase the budget of the principal.

1.1.2 Competing Incentives due to a Briber. In the second setting, we study competing incentives $C = (C_1, \dots, C_n)$ that are not intrinsic to the recommenders, but rather stem from bribes offered to recommenders by some third party. In our example, a recommender may not have direct ties to a construction company, but may instead be bribed by the construction company. Suppose the construction company gains a value $v \cdot A$, that is linear in the investment $A \geq 0$ into the flood barrier, where $v \in \mathbb{R}_+$ is the type of the briber and controls the amount by which it stands to gain. In this second setting, the briber may share part of the value they gain from the decision with recommenders, i.e., they may promise a recommender i a bribe $C_i \cdot A$, which the briber will pay to recommender i after the decision A has been taken. We have seen that we cannot guarantee that recommenders will not manipulate once there are competing incentives, but we may be able to dissuade a self-interested briber from choosing to bribe by making it too inefficient to bribe. We first give a necessary condition on the briber's value for the existence of mechanism that implements a target decision function f in this setting. Note that the game with the briber is a sequential game, and we use hence perfect Bayesian equilibria as solution concept, a refinement of Bayes-Nash.

Theorem 4. *(Informal): For any target decision function f, any mechanism MECH with budget $\beta > 0$, and any briber with value v, such that*

$$\frac{\beta}{v} \lessapprox \textsc{sensitivity}_f,$$

where $\textsc{sensitivity}_f$ is a measure of the sensitivity of f, there exist recommender beliefs Q such that all perfect Bayesian equilibrium reports R satisfy $F(R) \neq f(Q)$.

For the target decision function in Example 1, this translates to

$$\frac{\beta}{v} \lessapprox \frac{\ell^2}{n}.$$

Hence, if our budget is too low or the sensitivity of the target decision function (and hence also of the actual decision rule) is too large compared to the value a briber may gain from the decision, there will be a bribe and recommenders will not report truthfully. We again essentially match this bound with our adaptive-payment mechanism.

Theorem 5. *(Informal): For any target decision function f in some family of piecewise-linear functions,[2] any budget $\beta \geq 0$, and any briber value $v \geq 0$ such that*

$$\frac{\beta}{v} \gtrapprox \textsc{sensitivity}_f,$$

where $\textsc{sensitivity}_f$ is a measure of the sensitivity of f, the adaptive-payment mechanism ensures, for any beliefs Q, that all rationalizable report profiles are truthful and hence $F(R) = f(Q)$.

[2]For simplicity of exposition, we ignore in stating this theorem informally, a technicality that requires the target decision function to always take values > 0.

In particular, for the target decision function in Example 1, the condition in Theorem 5 translates to

$$\frac{\beta}{v} \gtrapprox \frac{\ell^2}{n}.$$

In the remainder of the paper, we will formally define our model and then state and prove these results.

1.2 Related work

The literature on elicitation with verification is built around *proper scoring rules*, which elicit subjective information from a single agent about an uncertain future event and align incentives with truthful reports [Brier et al., 1950, Gneiting and Raftery, 2007, Winkler, 1994]. A (strictly) proper scoring rule pays the agent a score based on their information and the event outcome; the score is (strictly) maximized when they truthfully-report their subjective information. *Wagering mechanisms* [Freeman and Pennock, 2018] and *prediction markets* [Wolfers and Zitzewitz, 2004] extend mechanisms for belief elicitation to multiple agents and allow for belief aggregation, but without modeling a decision to be made (and consequently, without handling competing incentives); see also Zhang et al. [2011].

In *decision scoring rules*, a principal uses agent reports to make a decision [Chen et al., 2011] and [York et al., 2021]. These papers consider the problem that arises when whether an uncertain event is observed depends on the decision. For example, if collecting information about the market success of a new electric vehicle (EV) design, with reports used to decide whether the EV is built and goes to market, then whether the outcome (market success or not) is observed depends on reports. In contrast to our work, Chen et al. [2011] and York et al. [2021] do not consider agents with outside incentives in regard to the decision that is made. To our knowledge, ours is the first paper in the information elicitation literature to formalize the problem of belief elicitation for the purpose of making a decision in the presence of participants with competing incentives.

The *peer prediction* literature [Jurca and Faltings, 2009, Miller et al., 2005, Prelec, 2004, Witkowski and Parkes, 2012, e.g.] studies *information elicitation without verification*—eliciting information from multiple agents in the absence of an uncertain future event against which to score. This absence of verifiable information is fundamental to the problem of peer prediction, and solutions rely on the correlation structure between agent signals. In this space as well, we are not aware of models that have handled the challenges that arise from the presence of competing incentives.

A body of work proposes strategy-proof mechanisms for cases when the mechanism designer and all agents know the distribution of valuations, and valuations are interdependent (e.g. Crémer and McLean [1988], Mezzetti [2004]). While these mechanisms are theoretically-interesting, they are difficult to practically implement because the principal frequently does not know either the distribution of valuations (beliefs) or the dependence of beliefs between agents. In our case, agents' beliefs and external incentives are independent, so the inference in the above work is not possible. Jehiel et al. [2006] show that ex-post efficient mechanisms are not possible in some settings while Bayesian "in expectation" efficient mechanisms are. Our work differs in that agents' predictions Q_i and their external incentives C_i are separate, while values in the above auction papers are the single item elicited.

Some work in social choice has studied a group decision problem with competing incentives, specifically selecting a committee from a set of voters where each voter would prefer to be selected [Alon et al., 2011]. They introduce a mechanism that aligns incentives through the use of randomization to introduce suitable independence between an agent's own report and whether or not it is selected; see also Kurokawa et al. [2015]. Another setting of competing incentives arises in the *transitive trust* literature, where reports on the trust one agent has for another are made, and a

mechanism is used to aggregate these reports. Competing incentives arise because a participant may care about their own trust ranking (e.g., being ranked above others, or above some threshold). Hopcroft and Sheldon [2007] develop a variation on PageRank [Page et al., 1999] that is robust to misreports; see also Parkes and Seuken [2022]. These settings are different from ours in that they are not about eliciting beliefs about an uncertain event, do not make use of payments, and the competing incentives stem from competition across participants for selection or ranking. Bribery has also been studied in voting settings [Elkind et al., 2009, Faliszewski and Rothe, 2016, Keller et al., 2018, Parkes et al., 2017], for example in regard to how many votes a briber must flip to change the outcome of an election or poll. In contrast, our setting treats the aggregation of belief reports and choosing an outcome in a continuous range, which is quite different from a discrete social choice problem; e.g., a participant with competing incentives may usefully be able to misreport by a small amount in our setting. The issue is that incentive to influence the decision is linear in decision probability, while derivative of proper score reduction asymptotically approaches zero at the true belief, meaning that there will be a small zone where the slope of outside incentive is greater than the slope of score incentive. We show how this can be handled in the setting where a rational briber is present.

The elements which make our problem setting unique include:

(1) The social choice functions of interest depend on payoff-independent signals, Q_i, not the external incentive C_i,
(2) direct-revelation mechanisms are precluded because the principal does not know the signal space Y_i (which may correspond to data), and subjective belief P_i is also similarly complex,
(3) the prior is unknown to the principal, and
(4) we have a contractible outcome.

2 PROBLEM DEFINITION

In this section, we formally define our model and discuss solution concepts.

2.1 Notation

Before we describe our model, we introduce some notation: We will use $[n]$ to denote the set $\{1, .., n\}$. If there is a variable x_i associated with each recommender $i \in [n]$, we use x to denote $(x_1, .., x_n)$ and if there is a set $\mathcal{X}_i$ associated with each recommender, we use $\mathcal{X}$ to denote the Cartesian product $\mathcal{X}_1 \times .. \times \mathcal{X}_n$. Further, we use the notation $\text{DISTR}(\mathcal{X}|\mathcal{Z})$ to denote the set of all conditional probability distributions over a random variable in $\mathcal{X}$ conditioned on a random variable in $\mathcal{Z}$. We use $\text{SUPP}(D)$ to denote the support of a probability distribution D.

2.2 Basic Set-up

There is a *principal*, whose seeks to elicit recommenders' beliefs about an unknown *outcome*, $O \in \mathcal{O} := \{0, 1\}$, in order to take an *action* $A \in \mathcal{A} := \mathbb{R}_{\geq 0}$. There are n *recommenders*, and each recommender $i \in [n]$ observes a *private signal* $Y_i \in \mathcal{Y}_i$ related to the outcome O (this signal, Y_i, could for example be some data that i has access to). Each recommender i has a *subjective belief* $P_i \in \mathcal{P}_i := \text{DISTR}(O|\mathcal{Y})$ regarding the outcome. While i will typically be uncertain about the signals of others, the conditional distribution P_i describes i's belief regarding O if i had access to the entire signal profile. This allows i to reason about the reports that others may make and how informed their reports may be. It also allows recommenders to hold different beliefs regarding O even if they had access to the same signal profile Y; i.e., it allows subjective beliefs, with different recommenders arriving at different conclusions even when presented with the same data.

Each recommender is asked to report their *subjective posterior belief* $Q_i \in \mathcal{Q}_i := [0,1]$ *regarding the probability of* $O = 1$, given their private signal, to a mechanism. As we shall see in the following sections, the belief Q_i follows from i's subjective belief $P_i(O|Y)$ a common prior regarding the private signals $D(Y)$, and the observed signal Y_i. Recommender i then submits a *report* $R_i \in \mathcal{R}_i := [0,1]$, which, if truthful, equals Q_i. A recommender may also have a *competing incentive*, $C_i \in \mathcal{C}_i := \mathbb{R}$, which may be positive or negative and yields value $C_i \cdot A$ to the recommender. This competing incentive is private, and distinct from any utility the recommender will gain as a result of receiving a payment in the mechanism, and may compete with the incentives provided by a mechanism for truthful reporting. We study two settings:

(1) In the first setting, each recommender may have an *intrinsic*, competing incentive regarding the outcome, which we model as $C_i \geq 0$.
(2) In the second setting, the competing incentive, C_i, represents a bribe offered by a self-interested briber. In this setting, the briber gains a value $V \cdot A$ that is proportional to the decision (e.g. the investment) where V is the type of the briber. The briber can then decide to share a fraction of their value with selected recommenders to incentivize them to misreport. The briber will determine these bribes such as to maximize the net value the briber gains from the decision, $(V - \sum_{i\in[n]} |C_i|) \cdot A$, while minimizing the payments they have to make to recommenders.

The goal of the principal is modeled through a *target decision function*, $f : \mathcal{Q} \to \mathcal{A}$, that expresses the action A that should ideally be taken based on recommenders' true, private beliefs Q. Since the principal does not have direct access to beliefs Q, the question is how to build a mechanism such that the actual action, A, is as close as possible to the target action, $f(Q)$.

2.3 An Elicitation Mechanism

To determine payments to recommenders and the action to be taken, the principal makes use of a *mechanism*, which is known to all agents and defined as follows:

Definition 1 (Mechanism). *A* mechanism *is defined by the tuple* $\text{MECH} = (F, \text{PAY})$*, where*

(1) $F : \mathcal{R} \to \mathcal{A}$ *is the mechanism's* decision rule*, which produces an action based on the reports, and*
(2) $\text{PAY} := (\text{PAY}_i)_{i\in[n]}$*, is the mechanism's* payment rule*, with* $\text{PAY}_i : \mathcal{R} \times \mathcal{O} \to \mathbb{R}_+$ *being a strictly proper scoring rule,*[3] *with respect to the report of* i*, that determines recommender* i*'s payment as a function of the reports and the observation.*

As discussed in the introduction, in the simplest case, the decision rule of the mechanism is simply the target decision function, i.e., $\mathrm{F} = f$ and hence $A = f(R)$. However, note that this will not necessarily yield the decision A closest to the target decision $f(Q)$, due to possible misreports $R \neq Q$. An F that differs slightly from f may (i) be less susceptible to misreports R, or (ii) it may even lead to better equilibria, where the reports R are closer to Q. Therefore, our mechanism design space allows for any F.

As discussed in Section 1.2, in the absence of competing incentives, using any proper scoring rule to determine recommenders' payments will ensure that our mechanism is truthful (dominant-strategy incentive compatible). It is hence natural to require that at least in that case, the mechanism should be truthful. Therefore Definition 1 only allows for strictly proper payment rules. Note, that

[3]Strictly proper requires $\mathbb{E}_{O\sim q_i}[\text{PAY}_i((q_i, r_{-i}), O)] > \mathbb{E}_{O\sim q_i}[\text{PAY}_i(r, O)]$ for all $\forall i \in [n], r \in \mathcal{R}, q_i \in \mathcal{Q}$. This requirement ensures that for honest recommenders, i.e. recommenders that have no competing incentives, the mechanism aligns incentives with truthful reporting. Many strictly proper scoring rules are known, see for instance Gneiting and Raftery [2007].

this imposes some restrictions on the form of $\text{PAY}_i(r, o)$ in r_i and o, but allows for any dependence in r_{-i}, including discontinuities.

For simplicity of exposition, we do not allow for randomization in our definition of a mechanism, but our results straightforwardly extend to this case, because our analysis only relies on the expected action and expected payments. We can hence alternatively also think of $\text{F}(R)$ as being the expected action, and $\text{PAY}_i(R, O)$ as the expected payment. For instance, we could interpret A as the probability of some binary decision (e.g., whether or not to build a particular flood barrier design).

2.4 The Elicitation Game with Intrinsic Competing Incentives

In the first setting, recommender i's type is (P_i, Y_i, C_i), and consists of i's subjective belief $P_i \in \mathcal{P}_i := \text{DISTR}(O|\mathcal{Y})$, private signal $Y_i \in \mathcal{Y}_i$, and competing incentive $C_i \in \mathcal{C}_i$. Agent types are private, and we model them as being drawn from a common prior, $D \in \text{DISTR}(\mathcal{P} \times \mathcal{Y} \times \mathcal{C})$, that is independent across recommenders and subtypes, i.e., prior D can be written as a product,

$$D(p, y, c) = \prod_{i\in[n]} D_i^p(p_i) \cdot \prod_{i\in[n]} D_i^y(y_i) \cdot \prod_{i\in[n]} D_i^c(c_i),$$

for some $D_i^p \in \text{DISTR}(\mathcal{P}_i)$, $D_i^y \in \text{DISTR}(\mathcal{Y}_i)$, $D_i^c \in \text{DISTR}(\mathcal{C}_i)$. Hence, the game played by recommenders in this setting is defined by the tuple (MECH, D), and proceeds as follows:

(1) The recommender types are drawn from the prior $P, Y, C \sim D$.
(2) Each recommender i observes their type (P_i, Y_i, C_i).
(3) Each recommender makes a report $R_i \in \mathcal{R}_i$.
(4) The mechanism takes action $A = \text{F}(R)$.
(5) The outcome $O \in \mathcal{O}$ is observed.
(6) Each recommender i receives utility $\text{PAY}_i(R, O) + C_i \cdot A$.

We model recommenders with utility linear in the payment received from the mechanism and their value $C_i \cdot A$ for the outcome. We do not specify, here, the distribution from which the outcome O is drawn. Rather, as discussed above, each recommender i holds a different subjective belief P_i regarding O. To formulate our game in the Bayesian setting, we express each recommender's subjective expected utility given the type profile p, y, c and report profile r. In fact, the subjective expected utility does not depend on the full type profile, but only on p_i, y, c_i, and we hence define:

$$\text{UTIL}_i\,(p_i, y, c_i, r) := \mathbb{E}_{O\sim p_i(\cdot|y)}\left[\text{PAY}_i(r, O) + c_i \cdot \text{F}(r)\right]. \tag{2}$$

This utility absorbs the random variable O by taking its expectation with respect to i's subjective belief. Hence, we will not need to reason about O explicitly in the sequel. Rather, we can now think of an equivalent version of the game, where steps (4) and (5) are replaced by each recommender i receiving utility $\text{UTIL}_i\,(P_i, Y, C_i, R)$.

2.4.1 Common Knowledge. We assume that the prior D and the mechanism MECH are known to all recommenders. However, we do not assume the principal knows the prior D and the principal cannot tailor their mechanism to the specific prior, we hence need a mechanism to yield good results for all priors D. In particular, the principal may not even know what kind of signal is available to recommenders, that is, the principal may not even know the space of observations $\mathcal{Y}$. For instance in Example 1 from the introduction, it is not plausible that one could model all the data different experts could possibly have regarding an extreme weather event.

2.4.2 Solution Concepts. Each recommender $i \in [n]$ will select their report R_i based on their type, (p_i, y_i, c_i), hence i's *strategy* has the form,

$$s_i \in \mathcal{S}_i := \text{DISTR}\left(\mathcal{R}_i^{\mathcal{P}_i\times C_i\times\mathcal{Y}_i}\right),$$

where we allow for mixed strategies. Recommender i's expected utility, given their type p_i, y_i, c_i and strategy profile s, is

$$\begin{aligned}\textsc{eutil}_i\left(s, p_i, y_i, c_i\right) &:= \mathbb{E}\left[\textsc{util}_i\left(P_i, Y, C_i, R\right) \middle| s, p_i, y_i, c_i\right] \\ &= \mathbb{E}_{P_{-i}, Y_{-i}, C_{-i} \sim D, \hat{S} \sim s}\left[\textsc{util}_i\left(p_i, \left(y_i, Y_{-i}\right), c_i, \hat{S}_i(p_i, y_i, c_i), \hat{S}_{-i}(P_{-i}, Y_{-i}, C_{-i})\right)\right]\end{aligned}$$

We model recommenders who play a strategy that is a best-response to some belief about other recommenders' strategies.

Definition 2. *Strategy $s_i \in \mathcal{S}_i$ is a* best-response *to a belief over others' strategies $\phi_i \in \mathit{distr}(\mathcal{S}_{-i})$ if*

$$\mathbb{E}_{S_{-i} \sim \phi_i}\left[\mathit{EUTIL}_i\left(\left(s_i, S_{-i}\right), p_i, y_i, c_i\right)\right] \geq \mathbb{E}_{S_{-i} \sim \phi_i}\left[\mathit{EUTIL}_i\left(\left(s_i', S_{-i}\right), p_i, y_i, c_i\right)\right] \quad \forall s_i' \in \mathcal{S}_i, p_i, y_i, c_i.$$

A natural solution concept in our setting is a Bayes-Nash equilibrium:

Definition 3 (Bayes-Nash Equilibrium (BNE):). *Strategy profile $s \in \mathcal{S}$ is a* BNE *if for all $i \in [n]$, s_i is a best response to the point belief $\phi_i \in \mathit{distr}(\mathcal{S}_{-i})$ that assigns probability 1 to s_{-i}.*

To strengthen our positive results, we define a second, more relaxed solution concept that captures a larger set of strategies that rational agents might play. Our notion is based on *rationalizability* [Bernheim, 1984, Pearce, 1984]. Informally, a strategy $s_i \in \mathcal{S}_i$ is rationalizable if it is a best-response to some belief $\phi_i \in \mathrm{distr}(\mathcal{S}_{-i})$ that i may hold about other agents' strategies. The key here is that ϕ_i cannot be arbitrary, it must be consistent with the knowledge that other agents are best-responding to their beliefs about other agents' strategies and that these beliefs are themselves also consistent with that knowledge etc. The set of rationalizable strategies is typically defined recursively: We start with the set of all strategy profiles, then, for each i, remove the strategies s_i that are not a best response to some belief over others' remaining strategies s_{-i} etc. until convergence, see e.g. [Shoham and Leyton-Brown, 2008]. For our purpose, it is sufficient to apply this recursion twice, which substantially simplifies the formal definition (and yields a superset of all rationalizable strategies, since each recursion only removes strategies):

Definition 4 (2-Rationalizability (2R):). *Let $\Phi := \textsc{distr}(\mathcal{S}_{-1}) \times .. \times \textsc{distr}(\mathcal{S}_{-n})$ be the set of belief profiles describing each agent's belief about other agents' strategies. We say that a strategy $s_i \in \mathcal{S}_i$ is 2R if there exists a belief $\pi_i \in \textsc{distr}(\mathcal{S}_{-i} \times \Phi_{-i})$ about other agents' strategies S_{-i} and beliefs ϕ_{-i} (recall, $\phi_j \in \textsc{distr}(\mathcal{S}_{-j})$ is j's belief regarding other agents' strategies), such that: (i) s_i is a best response to the marginal belief over strategies $\pi_i(S_{-i})$ and (ii) π_i is consistent with the knowledge that other agents will also play a best response strategy, i.e., for all $j \neq i$ and $\phi_j \in \Phi_j$, the conditional belief $\pi_i(S_j|\phi_j)$ only assigns nonzero probability to strategies S_j that are a best response to ϕ_j. We say that a strategy profile $s \in \mathcal{S}$ is 2R if all individual strategies $s_1, .., s_n$ are 2R.*

Hence, the set of 2R solutions contains the set of rationalizable solutions, which in turn contains the set of BNE solutions (see, e.g., [Shoham and Leyton-Brown, 2008]). Hence, our negative results, which hold for some BNE, also hold for a worst-case 2R profile. In contrast, we state our positive results for all 2R strategies, and hence they also hold for all BNE.

2.5 The Elicitation Game with a Briber

In the second setting, we introduce a briber who may choose to provide competing incentives to recommenders, when doing so is in the self interest of the briber. The elicitation game in the presence of a briber is similar to the one introduced in Section 2.4, except that the competing incentives are not generated by a prior D^c but generated instead by a rational agent (the briber). The solution concepts are very similar as well, and we relegate a detailed discussion of this setting to Section 4.

3 MANIPULATION DUE TO INTRINSIC COMPETING INCENTIVES

In this section, we quantify to what extent actual decision A may deviate from the target decision $f(Q)$ under different mechanisms. More precisely, we want to bound the difference $|A - f(Q)| = |\mathrm{F}(R) - f(Q)|$ between the action that is prescribed by the target decision function f for true beliefs Q, and the actual action $A = \mathrm{F}(R)$ taken by the mechanism based on the reported beliefs R under an equilibrium strategy. As discussed in Section 2.3, we do not limit the design space to mechanisms with decision rules that satisfy $\mathrm{F} = f$. For this reason, the design space $\textsc{mech} = (\mathrm{F}, \textsc{pay})$ is very large, as we allow for any arbitrary, even discontinuous F and PAY (as long as each $\textsc{pay}_i$ is strictly proper in r_i, o). For instance, we may hope to reduce or eliminate manipulation by co-designing F and PAY such that PAY incentivizes recommenders to report truthfully whenever they may have a large influence on F. Concretely, one could design F such that around any given report profile r only a small fraction of recommenders are influential and let PAY concentrate the entire budget on these recommenders.

Since there are so many possible designs, we start by giving a lower bound on the manipulation, relative to a given target decision function, that can be achieved by the best possible mechanism. We shall see that for any nontrivial target decision function and any mechanism with limited budget, there are, in the presence of nonzero competing incentives, equilibria with nonzero manipulation. The main task becomes quantifying how large this manipulation will be, depending on the target decision function f, the mechanism's budget β, the total competing incentive γ (i.e., the sum of competing incentives, formally defined below), and the number of recommenders n. We also present a novel mechanism that achieves this lower bound (up to log factors) for a family of piecewise-linear decision functions.

3.1 Lower Bounds on Manipulation

Intuitively, the more sensitive the target decision function f, the more attractive is manipulation for recommenders, as they can influence the decision with even small misreports (which incurs only a small loss in the payment due to the scoring rule PAY). In the following, we define a quantity we call Δ_f that describes the sensitivity of f. We will then give a lower bound on the manipulation as a function of Δ_f. In the next section, we shall see that this lower bound is tight for a family of target decision functions, which indicates that Δ_f is a fundamental property of f that captures its manipulability.

3.1.1 Sensitivity of Target Decision Functions. Our notion of sensitivity relies on a particular type of random walk. The idea is that this random walk moves through the domain of target decision function f, recommender by recommender:

Definition 5 (Random walk). *The random walk is defined by the tuple $\mathcal{W} = (\mathcal{K}, \mathcal{D}_0, \mathcal{D}_z)$ where $\mathcal{K} \subseteq [n]$ is a subset of recommenders, and $\mathcal{D}_0, \mathcal{D}_z$ are distributions with support in $[0,1]^n$ such that $X^0 \sim \mathcal{D}_0$, $Z \sim \mathcal{D}_z$ satisfy $X^0 + Z \in [0,1]^n$ with probability 1. We define a random walk to be the following process: First, $\Theta^1, .., \Theta^k \in \mathcal{K}$, are drawn uniformly without replacement from $\mathcal{K}$. Then, $X^1, .., X^k$ are generated as follows*

$$X^0 \sim \mathcal{D}_0; \quad Z \sim \mathcal{D}_z; \quad X_i^j = \begin{cases} X_i^{j-1} + Z_i & \text{if } i = \Theta^j \\ X_i^{j-1} & \text{else.} \end{cases}$$

Let $\textsc{walks}(k, \delta)$ be the set of all such random walks $\mathcal{W} = (\mathcal{K}, \mathcal{D}_0, \mathcal{D}_z)$ with $|\mathcal{K}| = k$ and $\mathcal{D}_z$ having support $[-\delta, \delta]^n$.

Given this definition of a random walk, we can now define sensitivity.

Definition 6 (Sensitivity Δ_f). *For a given target decision function f, we define*

$$\Lambda_f(k,\delta) := \max_{\mathcal{W}\in \mathit{WALKS}(k,\delta)} \mathbb{E}_{X\sim\mathcal{W}}\left[\sum_{i=1}^{k} \left|f(X^i) - f(X^{i-1})\right|\right],$$

and

$$\Delta_f := \max_{k\in[n],\delta\in[0,1]} \frac{\Lambda_f(k,\delta)^2}{\delta\cdot k\cdot \log(3\cdot k)}.$$

We will call Λ_f the sensitivity function and Δ_f the sensitivity.

Intuitively, the sensitivity function $\Lambda_f(k,\delta)$ corresponds to the sum of variations of the target decision function f along a random walk through its domain, involving k recommenders, each of whom vary their report/belief by at most γ. We will see how to construct a random walk and obtain $\Lambda_f(k,\delta)$ for some families of target decision functions in the next section. But first, we'll show how this notion of sensitivity translates into a lower bound on manipulation.

3.1.2 General Lower Bound on Manipulation. We provide a lower bound on the extent to which the actual decision A may deviate from the target decision $f(Q)$ under any mechanism MECH (and hence even under the optimal mechanism for f). This lower bound depends on the sensitivity Δ_f of the target function, the number of recommenders n, the budget β, and the total competing incentive γ, which we define as $\gamma := \max_{c\in\mathrm{SUPP}(D^c)} \|c\|_1$.

Theorem 6 (Lower-bound on multi-recommender manipulability)**.** *Suppose we want to implement some target decision function f. Then, for any $\beta,\gamma\geq 0$ and any mechanism MECH with budget β, there is a prior D with total competing incentive γ such that for all BNE we have with probability* 1

$$|F(R) - f(Q)| \geq \frac{1}{16}\cdot \max_{k\in[n],\delta\in[0,1]} \frac{\Lambda_f(k,\delta)^2}{k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot\log(3\cdot k) + k\cdot\Lambda_f(k,\delta)}.$$

For small competing incentives γ, we show that this implies

$$|F(R) - f(Q)| \geq \frac{1}{16}\cdot\frac{\gamma}{\beta}\cdot\Delta_f - O(\gamma^2) \quad \text{as } \gamma\to 0.$$

Even with an arbitrarily small competing incentive $\gamma > 0$, no mechanism can prevent manipulation, unless it ignores the recommenders' reports, i.e., $\Lambda_f(k,\delta) = 0$. Thus, even if we co-design the decision F and payment PAY function in some intricate way, allowing even for discontinuities, it is impossible to guarantee that there will be no manipulation. Further, this result implies that any target decision function f with a discontinuity of size d (which implies $\Lambda_f(1,0) = d$) has manipulability of at least $\frac{d}{16}$ for any $\gamma > 0$ and any budget β. Finally, we see that, for small competing incentive γ, the manipulation scales linearly with γ/β (the ratio between the competing incentive and the budget of the mechanism) and with the sensitivity Δ_f of f. We see that the quantity Δ_f captures the vulnerability of target decision function f to manipulation, and sensitivity will play a central role in the sequel. We shall now see how to express this lower bound for particular classes of target decision functions.

3.1.3 Lower Bounds on Manipulation for Target Decision Functions with Bounds on the Derivative. It is instructive to specialize Theorem 6 to the case where, in some part of its domain, the target decision function f is known to have a lower bound on its derivative.[4] The proof of the following result also illustrates how a random walk can be constructed in order to find the sensitivity Δ_f of a given f.

[4]This does not impose any restrictions on the design space of the mechanism, i.e., the decision rule is still allowed to be arbitrary, even discontinuous.

Lemma 1. *Consider a target decision function f with the following property: There is a set of recommenders $\mathcal{M} \subseteq [n]$, a subset of the domain $\mathcal{L} \subseteq \mathcal{Q}$, and an $L \in \mathbb{R}_+$, such that for each recommender $i \in \mathcal{M}$, the decision function is monotonic and has slope at least L in $\mathcal{L}$, i.e.*

$$\frac{\partial f}{\partial x_i}(x) \geq L \quad x \in \mathcal{L}, \quad \text{or} \quad \frac{\partial f}{\partial x_i}(x) \leq -L \quad x \in \mathcal{L}$$

where $\mathcal{L}$ is a hypercube, $\mathcal{L} = \left\{x \in \mathbb{R}^n : x_{-\mathcal{M}} = \mu_{-\mathcal{M}}, \|x_{\mathcal{M}} - \mu_{\mathcal{M}}\|_\infty \leq \sigma\right\}$, parametrized by $\sigma \in [0,1]$, and $\mu \in [\sigma, 1-\sigma]^n$. For any such target decision function f, we have (using $m := |\mathcal{M}|$)

$$\Delta_f \geq \frac{m}{\log(3 \cdot m)} \cdot \sigma \cdot L^2.$$

Proof. For a given $k \leq m, \delta \leq \sigma$, consider the random walk with $\mathcal{K} \subseteq \mathcal{M}$ some set of size k, $X^0 = \mu$, and $Z_i = \delta$ for $i \in \mathcal{K}$ and $Z_i = 0$ else. Then, for any such path, we have

$$\sum_{j=1}^{k} \left| f(X^j) - f(X^{j-1}) \right| \geq \delta \cdot L \cdot k.$$

Hence, for any $k \leq m$ and $\delta \leq \sigma$, we have $\Lambda_f(k, \delta) \geq \delta \cdot L \cdot k$. It then follows that,

$$\begin{aligned}\Delta_f &= \max_{k \in [n], \delta \in [0,1]} \frac{\Lambda_f(k,\delta)^2}{k \cdot \delta \cdot \log(3 \cdot k)} \geq \max_{k \in [m], \delta \leq \sigma} \frac{\Lambda_f(k,\delta)^2}{k \cdot \delta \cdot \log(3 \cdot k)} \geq \max_{k \in [m], \delta \leq \sigma} \frac{\delta \cdot L^2 \cdot k}{\log(3 \cdot k)} \\ &\geq \frac{\sigma \cdot L^2 \cdot m}{\log(3 \cdot m)} \quad (\text{with } \delta = \sigma, k = m).\end{aligned}$$

□

Combining this result with Theorem 6, we immediately obtain that, under the conditions of Theorem 6, we have

$$|\mathrm{F}(R) - f(Q)| \geq \frac{1}{16} \cdot \frac{\gamma}{\beta} \cdot \frac{m}{\log(3 \cdot m)} \cdot \sigma \cdot L^2 - O(\gamma^2).$$

We see that the larger the set $\mathcal{M}$ of recommenders with respect to whose reports the slope of the target decision function is large, the larger the manipulation. Further, the manipulation also grows with the size σ of the domain within which these recommenders have a large influence. Finally, the manipulation grows quadratically with the slope L of f. Using a more sophisticated random walk, we can obtain the following more general result:

Lemma 2. *Consider a decision function f with the following property: There is a set of recommenders $\mathcal{M} \subseteq [n]$, a subset of the domain $\mathcal{L} \subseteq \mathcal{Q}$, and an $L \in \mathbb{R}_+$, such that for each recommender $i \in \mathcal{M}$, the decision function is monotonic and has slope at least L in $\mathcal{L}$, i.e.*

$$\frac{\partial f}{\partial x_i}(x) \geq L \quad x \in \mathcal{L}, \quad \text{or} \quad \frac{\partial f}{\partial x_i}(x) \leq -L \quad x \in \mathcal{L},$$

where $\mathcal{L}$ is the set

$$\mathcal{L} := \left\{ x \in \mathbb{R}^n : x_{-\mathcal{M}} = \mu_{-\mathcal{M}}, \|x_{\mathcal{M}} - \mu_{\mathcal{M}}\|_\infty \leq \sigma, \left| \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} (x_i - \mu_i) \right| \leq \epsilon \right\},$$

and parametrized by $\sigma \in [0,1], \mu \in [\sigma, 1-\sigma]^n$, and $\epsilon \in [0,1]$. For any target decision function f with this property, we have, using $m := |\mathcal{M}|$,

$$\Delta_f \geq \frac{1}{3} \cdot \frac{m}{\log(3 \cdot m)} \cdot \sigma \cdot L^2 \cdot \min\left\{ \frac{\sqrt{m} \cdot \epsilon}{\sigma}, 1 \right\}.$$

We will study the implications of this result for specific decision functions in Section 3.3. But first, we give a novel mechanism and prove an upper bound on its degree of manipulation. The negative results from the present section are helpful for designing mechanisms, since they serve to give a bound on what is possible.

3.2 The Adaptive-Payment Mechanism

We start by defining a second, more conservative notion of sensitivity that will be helpful for our upper bounds.

Definition 7 (Sensitivity ∇_f). *For a target decision function f that is differentiable almost everywhere, we define another sensitivity notion,*

$$\nabla_f := \left(m \cdot \int_0^1 \max_{i\in[n], q_{-i}\in Q_{-i}} \left|\partial_{q_i} f(q)\right| dq_i \right) \cdot \left(\max_{i\in[n], q\in Q} \left|\partial_{q_i} f(q)\right| \right),$$

where m is the number of recommenders that f has nonzero dependence on, and ∂ denotes the largest generalized derivative (identical to the derivative wherever f is differentiable).

We shall see in Section 3.3 that ∇_f is equivalent to Δ_f (up to log factors) for a family of piecewise-linear target decision functions. We now give a novel mechanism, that we show achieves the lower bounds from Section 3.1 for a particular class of target decision functions. The key idea is to define the payment rule such that it concentrates the incentives where the target decision function f has a large derivative.

Definition 8 (Adaptive-payment mechanism). *Given a budget β and a target decision function f that is differentiable almost everywhere, we define the* adaptive payment rule, *(f,β)-ada-pay, as follows: Let $\mathcal{M} \subseteq [n]$ be the set of recommenders on which f has nonzero dependence. Then, let*

$$\textit{ADA-PAY}_i(r_i, o) = \begin{cases} \frac{\beta}{3\cdot|\mathcal{M}|} \cdot \left(\frac{\int_0^{r_i} \delta_i(x)(o-x)\cdot dx}{\int_0^1 \delta_i(x)\cdot dx} + 1 \right) & \text{if } i \in \mathcal{M} \\ \frac{\beta}{6\cdot n} \cdot \left(\int_0^{r_i} (o-x)\cdot dx + 1 \right) & \text{else} \end{cases}$$

with $\delta_i : [0,1] \to \mathbb{R}_{>0}$ being a continuous function such that

$$\delta_i(r_i) = \max_{r_{-i}} \left|\partial_{r_i} f(r)\right| + \int_0^1 \max_{r_{-i}} \left|\partial_{r_i} f(r)\right| dr_i,$$

where ∂ denotes the largest generalized derivative (just the derivative wherever f is differentiable). We call $\textit{MECH} = (F = f, \textit{PAY} = (f,\beta)\text{-ADA-PAY})$, the adaptive-payment mechanism for decision function f and budget β.

To gain an intuition for this payment rule, consider a single recommender with true belief q. The derivative of the expected payment, under adaptive payment rule, has the form

$$\begin{aligned} \partial_r \mathbb{E}_{o\sim q}\left[\textsc{ada-pay}(r,o)\right] &= \partial_r \left(\frac{\beta}{3} \cdot \frac{\int_0^r \delta(x)(q-x)\cdot dx}{\int_0^1 \delta(x)\cdot dx} \right) \\ &= \frac{\beta}{3} \cdot \frac{\delta(r)(q-r)}{\int_0^1 \delta(x)\cdot dx} \\ &= \frac{\beta}{6} \cdot \left(1 + \frac{|\partial_r f(r)|}{\int_0^1 |\partial_x f(x)|\, dx} \right)(q-r). \end{aligned}$$

First of all, observe that the derivative is zero only if $q = r$, hence the rule is proper. The novelty of this payment rule is that its derivative scales with $|\partial_r f(r)|$, i.e., it concentrates incentives where misreports have a large influence on the decision.

Using the adaptive payment rule, we obtain a mechanism that achieves the following upper bound on manipulation:

Theorem 7. *Suppose that we are given target decision function f that is differentiable almost everywhere. Using the adaptive-payment mechanism Definition 8 with budget β, we have for any $\gamma \geq 0$, any prior D with total competing incentive γ, and any 2R strategy profile,*

$$|F(R) - f(Q)| \leq 6 \cdot \frac{\gamma}{\beta} \cdot \nabla_f,$$

with probability 1.

This dependence on γ and β matches the one in the lower bound (Theorem 6). To gain further understanding of this result, we will focus on a particular class of decision functions.

Lemma 3. *Consider a target decision function f with the following properties: It is differentiable almost everywhere, there is some set of recommenders $\mathcal{M} \subseteq [n]$ of size m and some parameters $\mu \in \mathcal{R}, \sigma \in [0, 1]$ (with $\mu \pm \sigma \in \mathcal{R}$) such that the derivative of f is upper bounded by*

$$\left|\frac{\partial f}{\partial x_i}(x)\right| \leq \begin{cases} L & \text{if } i \in \mathcal{M} \text{ and } x \in \mathcal{L}, \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{L} := \left\{x \in \mathbb{R}^n : \|x - \mu\|_\infty \leq \sigma\right\}$ is a hypercube with center μ and side length σ. For any such decision function, we have

$$\nabla_f \leq 2 \cdot m \cdot \sigma \cdot L^2.$$

Proof. Such a target decision function f satisfies for all i

$$\int_0^1 \max_{r_{-i}} |\partial_x f(x, r_{-i})| \, dx \leq 2 \cdot \sigma \cdot L \quad \text{and} \quad \max_r \left|\partial_{r_i} f(r)\right| \leq L.$$

Further, by definition, f only depends on m recommenders. Hence, it follows immediately from Definition 7 that

$$\nabla_f \leq (m \cdot 2 \cdot \sigma \cdot L) \cdot (L) = 2 \cdot \sigma \cdot m \cdot L^2.$$

□

It follows immediately from Lemma 3 and Theorem 7 that, under the conditions of Theorem 7, we have

$$|\mathrm{F}(R) - f(Q)| \leq 12 \cdot \frac{\gamma}{\beta} \cdot m \cdot \sigma \cdot L^2.$$

This bound matches the lower bound below Lemma 1 up to log factors. Hence, for decision functions that are linear on some hypercube $\mathcal{L}$ and constant elsewhere, our mechanism is essentially optimal and the lower and upper bounds are tight.

3.3 Bounded-Linear Target Decision Functions

We will now define a family of piecewise-linear target decision functions, which will allow us to illustrate and compare the general results we obtained in the previous sections. For convenience, we will use the notation $[\![x]\!]_y^z$ (for some $x, z \geq y \in \mathbb{R}$) to denote the value of x clipped between y and z, i.e. $[\![x]\!]_y^z := \min\{\max\{x, y\}, z\}$.

Definition 9 (Bounded-linear target decision function). *We say that a target decision function f is bounded-linear if it can be written as*

$$f(q) = \bar{f} + \left[\!\!\left[\ell \cdot \frac{1}{m} \sum_{i \in \mathcal{M}} [\![q_i - t]\!]_{-\alpha}^{\alpha} \right]\!\!\right]_{-\varepsilon}^{\varepsilon}$$

with parameters $\mathcal{M} \subseteq [n]$ (we use $m := |\mathcal{M}|$), $\bar{f} \in \mathbb{R}$, $t \in \left[\frac{1}{10}, \frac{9}{10}\right]$, $\alpha \in [0, 1]$, $|\ell| \leq \frac{\sqrt{m}}{\alpha}$, $\varepsilon \geq \frac{1}{10} \cdot \alpha \cdot \ell$. Hence, target decision function f is parametrized by $(\mathcal{M}, \bar{f}, \varepsilon, \ell, \alpha, t)$.

For instance, with $\mathcal{M} = [n]$ and $\alpha = 1$, $\varepsilon = \infty$, we obtain a target decision function that is linear in the average report $\bar{q} = \sum_{i \in [n]} q_i$,

$$f(q) = \bar{f} + \ell \cdot (\bar{q} - t).$$

We shall see other instances of useful bounded-linear decision functions below, but first we derive some general results. Recall that we introduced two notions of sensitivity, Δ_f and ∇_f. We used Δ_f in our lower bounds on manipulation and ∇_f in our upper bounds. We shall now see that for bounded-linear decision functions, the two notions are equivalent up to log factors and hence our bounds are tight.

Lemma 4. *For any bounded-linear target decision function, f, as defined in Definition 9, we have*

$$\nabla_f \leq 2 \cdot \frac{\alpha \cdot \ell^2}{m} \leq 60 \cdot \log(3 \cdot m) \cdot \Delta_f.$$

Proof. We first proof an upper bound on ∇_f and then a lower bound on Δ_f, and then we relate them: It follows straightforwardly from Lemma 3, with $L = \frac{\ell}{m}$ and $\sigma = \alpha$, that

$$\nabla_f \leq 2 \cdot \frac{\alpha \cdot \ell^2}{m}. \tag{3}$$

We now proceed to prove a lower bound on Δ_f. To apply Lemma 2, note that any $q \in \mathcal{Q}$ such that

$$\|q_{\mathcal{M}} - t\|_\infty \leq \min\{t, 1 - t, \alpha\} \quad \text{and} \quad \left| \frac{1}{m} \sum_{i \in \mathcal{M}} q_i - t \right| \leq \varepsilon / \ell$$

we have $\frac{\partial f}{\partial q_i}(q) = \frac{\ell}{m}$. Further, it follows from Definition 9 that $\min\{t, 1 - t, \alpha\} \leq \frac{\alpha}{10}$, and we can hence apply Lemma 2, with $\sigma \to \frac{\alpha}{10}, L \to \frac{\ell}{m}, \epsilon \to \varepsilon/\ell$, and obtain

$$\Delta_f \geq \frac{1}{3} \cdot \frac{m}{\log(3 \cdot m)} \cdot \sigma \cdot L^2 \cdot \min\left\{ \frac{\sqrt{m} \cdot \epsilon}{\sigma}, 1 \right\} = \frac{1}{30} \cdot \frac{1}{\log(3 \cdot m)} \cdot \frac{\alpha \cdot \ell^2}{m} \cdot \min\left\{ 10 \cdot \frac{\varepsilon \cdot \sqrt{m}}{\alpha \cdot \ell}, 1 \right\},$$

and since Definition 9 specifies that $\varepsilon \geq \frac{1}{10} \cdot \alpha \cdot \ell$, we have

$$\Delta_f \geq \frac{1}{30} \cdot \frac{1}{\log(3 \cdot m)} \cdot \frac{\alpha \cdot \ell^2}{m}.$$

The statement in the lemma follows straightforwardly by comparing this inequality with (3). □

We can now state both the lower bound on manipulation (Theorem 6) and the upper bound (Theorem 7) in terms of Δ_f.

Theorem 8. *Suppose we want to implement some bounded-linear target decision function f (Definition 9). Then, for any mechanism $\text{MECH} = (F, \text{PAY})$ with budget β and any total competing incentive γ, there is a prior D such that for all BNE, with probability* 1*, the true beliefs Q and the report R are such that*

$$|F(R) - f(Q)| \geq \frac{1}{16} \cdot \frac{\gamma}{\beta} \cdot \Delta_f - O(\gamma^2)$$

$$\text{(with } m, \alpha, \ell \text{ being params of } f\text{, see Definition 9)} \geq \frac{2}{10^3 \cdot \log(3 \cdot m)} \cdot \frac{\gamma}{\beta} \cdot \frac{\alpha \cdot \ell^2}{m} - O(\gamma^2)$$

$$\geq \frac{1}{10^3 \cdot \log(3 \cdot m)} \cdot \frac{\gamma}{\beta} \cdot \nabla_f - O(\gamma^2) \quad \text{as } \gamma \to 0$$

Using the adaptive-payment mechanism (Definition 8) with budget β, we have for any $\gamma \geq 0$, any prior D with total competing incentive γ, and any 2R strategy profile,

$$|F(R) - f(Q)| \leq 6 \cdot \frac{\gamma}{\beta} \cdot \nabla_f \leq 12 \cdot \frac{\gamma}{\beta} \cdot \frac{\alpha \cdot \ell^2}{m} \leq 360 \cdot \log(3 \cdot m) \cdot \frac{\gamma}{\beta} \cdot \Delta_f$$

with probability 1*.*

The two bounds are equivalent, up to a log factor. Hence for the family of bounded-linear target decision functions, the quantities Δ_f and ∇_f are essentially equivalent, and each one of them fully captures how sensitive f is to manipulation. Further, Theorem 8 implies that the adaptive payment mechanism (Definition 8) is optimal (up to log factors) for bounded-linear decision functions. We see that the dependence of the manipulation on the parameters of the decision function is $\frac{\alpha \cdot \ell^2}{m}$. With $\alpha = 1$, we obtain a linear target decision function in the mean estimator, $\bar{q}$, i.e.,

$$f(q) = \bar{f} + [\![\ell \cdot \bar{q} - t]\!]_{-\varepsilon}^{\varepsilon}$$

with manipulation $\frac{\ell^2}{m}$. We can interpret an $\alpha < 1$ as a more robust estimator, where we limit the influence of each recommender to α, and obtain a better manipulation bound $\frac{\alpha \cdot \ell^2}{m}$.

4 MANIPULATION DUE TO BRIBERY

In this section, we study the second setting, where the competing incentives are not intrinsic to the recommenders but stem from a rational third party that we call the briber. We shall see that in this setting, manipulation can be avoided entirely. We will give necessary and sufficient conditions for avoiding manipulation in terms of the sensitivity of the decision function f, the strength of the incentive of the briber, and the budget of the mechanism.

4.1 Setting

We now adapt the setting for intrinsic competing incentives defined in Section 2.4 to incentives stemming from a rational briber. The *briber's type* (V, W) consists of the value $V \in \mathcal{V} := \mathbb{R}_+$ the briber gains from an increase in the decision A (in our example, the construction company gaining value from investments into flood barriers), and let $W \in \mathcal{W} := 2^{[n]}$ denote the set of recommenders the briber can access. The competing incentive C is now determined by the briber and hence not part of the recommenders' types anymore. Each *recommender's type* (P_i, Y_i) now consists of the subjective belief $P_i \in \mathcal{P}_i := \text{DISTR}(O|\mathcal{Y})$ and private signal $Y_i \in \mathcal{Y}_i$. Agent types are private, and we

model them as being drawn from a common prior, $D \in \text{DISTR}(\mathcal{P} \times \mathcal{Y} \times \mathcal{V} \times W)$, that is independent across agents and subtypes, i.e., prior D can be written as a product,

$$D(p, y, v, w) = \prod_{i \in [n]} D_i^p(p_i) \cdot \prod_{i \in [n]} D_i^y(y_i) \cdot D^v(v) \cdot D^w(w),$$

for some $D_i^p \in \text{DISTR}(\mathcal{P}_i)$, $D_i^y \in \text{DISTR}(\mathcal{Y}_i)$, $D^v \in \text{DISTR}(\mathcal{V})$, $D^w \in \text{DISTR}(\mathcal{W})$. In the presence of a briber, the elicitation game proceeds as follows:

(1) The types are drawn from the prior $P, Y, V, W \sim D$.
(2) Each recommender i observes their type (P_i, Y_i) and the briber observes their type V, W.
(3) *The briber determines the bribe profile $C \in \mathcal{C}$ (note that it must satisfy $C_i = 0$ for all $i \notin W$).*
(4) Each recommender i observes their bribe C_i.
(5) Each recommender makes a report $R_i \in \mathcal{R}_i$.
(6) The mechanism takes action $A = \text{F}(R)$,
(7) The outcome $O \in \mathcal{O}$ is observed.
(8) Each recommender i receives utility $U_i = \text{PAY}_i(R, O) + C_i \cdot A$.
(9) *The briber receives utility* $U^{BRI} := \left(V - \sum_{i \in [n]} |C_i|\right) \cdot A$.

As for the game with intrinsic competing incentives, we formulate our modified game in the Bayesian setting, hence we express each agent's subjective expected utility given the type profile p, y, v, w and action profile r, c. Recommender i's subjective expected utility only depends on p_i, y, c_i, r:

$$\text{UTIL}_i\left(p_i, y, c_i, r\right) := \mathbb{E}_{O \sim p_i(\cdot|y)}\left[\text{PAY}_i(r, O) + c_i \cdot \text{F}(r)\right] \tag{4}$$

which is identical to the recommenders' utility stated in the setting with intrinsic competing incentives (2), except that c_i now represents the briber's action, rather than i's own type.

The briber's utility only depends on briber's own type, v, w, the bribe profile c, and the report profile r

$$\text{UTIL}^{\text{BRI}}\left(v, w, c, r\right) := \left(v - \sum_{i \in [n]} |c_i|\right) \cdot \text{F}(r).$$

4.1.1 Common Knowledge. We assume that the prior D and the mechanism MECH are known to all recommenders and the briber. As before, we do not assume that the principal knows the prior D.

4.1.2 Solution Concepts. For recommenders, the set of possible strategies $\mathcal{S}$ is the same as in the game without briber, i.e.

$$s_i \in \mathcal{S}_i := \text{DISTR}\left(\mathcal{R}_i^{\mathcal{P}_i \times C_i \times \mathcal{Y}_i}\right).$$

For the briber, the strategy has the form

$$s^{\text{BRI}} \in \mathcal{S}^{\text{BRI}} := \text{DISTR}\left(\mathcal{C}^{\mathcal{V} \times \mathcal{W}}\right),$$

and we will use s_i^{BRI} to denote the bribe paid to recommender i. The briber's expected utility, given the briber type and a strategy profile, is

$$\begin{aligned}\text{EUTIL}^{\text{BRI}}\left(s, s^{\text{BRI}}, v, w\right) &:= \mathbb{E}\left[\text{UTIL}^{\text{BRI}}\left(V, W, C, R\right) \middle| s, s^{\text{BRI}}, v, w\right] \\ &= \mathbb{E}_{P,Y,V,W \sim D, \hat{S} \sim s, \hat{S}^{\text{BRI}} \sim s^{\text{BRI}}}\left[\text{UTIL}^{\text{BRI}}\left(V, W, \hat{S}^{\text{BRI}}(V, W), \hat{S}\left(P, Y, \hat{S}^{\text{BRI}}(V, W)\right)\right)\right].\end{aligned}$$

Recommender i's expected utility given their type p_i, y_i, bribe c_i (which they know at the time of reporting), and a strategy profile, is

$$\text{EUTIL}_i\left(s, s^{\text{BRI}}, p_i, y_i, c_i\right) := \mathbb{E}\left[\text{UTIL}_i\left(P_i, Y, C_i, R\right) \middle| s, s^{\text{BRI}}, p_i, y_i, c_i\right].$$

We express this expectation in full in Appendix A. As a solution concept in this sequential move game between the briber and the recommenders, we adopt the notion of a *perfect Bayesian equilibrium*, where a belief is said to be consistent if it follows from Bayes rule, whenever it is well-defined, and in ill-defined cases (when the event we condition on has zero probability, as happens when c_i is not in the support of the mixed strategy of the briber s^{BRI}), any belief is said to be consistent. We discuss the technical details of what consistency implies regarding recommenders' beliefs in Appendix A.

Definition 10 (Perfect Bayesian equilibrium (PBE)). *A strategy profile s, s^{BRI} is a* perfect Bayesian equilibrium *if*

$$EUTIL^{BRI}\left(s, s^{BRI}, v, w\right) \geq EUTIL^{BRI}\left(s, s'^{BRI}, v, w\right) \quad \forall s'^{BRI}, v, w$$

and for each recommender i, there is a consistent belief such that the implied expected utility satisfies

$$EUTIL_i\left(s, s^{BRI}, p_i, y_i, c_i\right) \geq EUTIL_i\left(s_i', s_{-i}, s^{BRI}, p_i, y_i, c_i\right) \quad \forall s_i', p_i, y_i, c_i.$$

For our rationalizability notion, 2R, we do not need to introduce such a refinement, as we only use it for positive results, which are not weakened by including non-credible strategy profiles.

4.2 Results: Elicitation in the Presence of a Briber

We know that whenever the competing incentive of the recommender is non-zero and the decision function is nontrivial, then no mechanism is truthful. But what if the competing incentive comes from a rational briber? In this case we recover a positive result, by using a mechanism in which it is not in the interest of the briber to make a bribe. We characterize the conditions under which a target decision function f can be implemented, by providing a necessary and then a sufficient condition on the sensitivity of f, the mechanism's budget β, and the briber incentive v, which we define as $v := \max_{v \in \text{SUPP}(D^v)} v$.

Theorem 9. *Suppose we want to implement some target decision function f with sensitivity Δ_f (Definition 6). In the briber setting with incentive v, for any mechanism MECH with budget β such that*

$$\frac{\beta}{v} \leq \frac{1}{8 \cdot \max_{q \in Q} f\left(q\right)} \cdot \Delta_f$$

there is a prior profile D such that for all PBE (Definition 10), the action taken is not identical to the target action, i.e. $F(R) \neq f(Q)$, with probability 1.

We also provide the following sufficient condition.

Theorem 10. *Suppose we want to implement some target decision function f with sensitivity ∇_f (Definition 7). In the briber setting, if we use the adaptive-payment mechanism (Definition 8) with budget β, then, for any prior D with briber incentive v such that*

$$\frac{\beta}{v} \geq \frac{6}{\min_{q \in Q} f\left(q\right)} \cdot \nabla_f,$$

all 2R strategy profiles (Definition 4) are such that, with probability 1*, the action taken based on the reports, $F(R)$, is identical to the target action $f(Q)$.*

Thus, we see that in the briber setting, it is possible to ensure that all rationalizable strategies are truthful. This is possible if and only if the budget of the principal, β, is proportional to the incentive of the briber, v. To gain further intuition into the dependence on the sensitivity of the

target decision function f, we can apply these results to the family of bounded-linear target decision functions. From Definition 9, it follows immediately that the necessary condition translates to

$$\frac{\beta}{\upsilon} \leq \frac{2}{10^3 \cdot \log(3 \cdot m) \cdot \max_{q \in Q} f(q)} \cdot \frac{\alpha \cdot \ell^2}{m},$$

and the sufficient condition to

$$\frac{\beta}{\upsilon} \geq \frac{12}{\min_{q \in Q} f(q)} \cdot \frac{\alpha \cdot \ell^2}{m}.$$

These two conditions match up to log factors and the factor $\max_{q \in Q} f(q) / \min_{q \in Q} f(q)$. Assuming that this ratio is bounded, the key dependence of $\frac{\beta}{\upsilon} \propto \frac{\alpha \cdot \ell^2}{m}$ appears in both bounds, which establishes it as a necessary and sufficient condition for the existence of a mechanism that ensures that only truthful strategies are rationalizable. It is noteworthy that the factor $\frac{\alpha \cdot \ell^2}{m}$ also determined the amount of manipulation in the setting of intrinsic competing incentives, see Theorem 8. Hence, this factor appears to capture, in a very general sense, the sensitivity to competing incentives of bounded-linear decision function.

A numeric example: To gain intuition into the magnitude of the budget β required to avoid manipulation, we consider a numeric example. A city plans to invest between 30 and 130 million dollars into flood defenses, proportional to the estimated probability $\bar{Q} = \frac{1}{m} \sum_i Q_i$ of an extreme weather event:

$$f(Q) = 3 \cdot 10^7 + 10^8 \cdot \bar{Q}.$$

A construction company gains a profit of 5% of the investment (i.e. $\upsilon = 0.05$) and therefore bribes recommenders. Combining Theorem 10 with Lemma 3, it follows that a total budget of $\beta \geq 10^8 / m$ guarantees truthful reports. Hence, with a single recommender we could guarantee truthful reporting with payments of 100 million dollars, which is clearly not feasible. However, with 20 recommenders, 5 million would be sufficient, and with 100 recommenders even 1 million would be sufficient.

5 CONCLUSION

We have characterized the extent to which recommenders will manipulate a decision that a principal wants to make according to a target decision function f. The key quantities influencing the magnitude of the manipulation are (i) the strength of recommenders' or the briber's competing incentives, γ, (ii) the budget of the principal, β, and (iii) the sensitivity Δ_f of the target decision function f. The quantity that has appeared repeatedly throughout the paper and characterizes to what extent the decision is manipulated is $\frac{\gamma}{\beta} \cdot \Delta_f$. For the class of bounded-linear target decision functions, this quantity determines the upper and the lower bound for manipulation, indicating that our notion of sensitivity captures a fundamental property of the decision function. For this class of target decision functions, the sensitivity takes the form $\frac{\alpha \cdot \ell^2}{n}$, where ℓ is the slope, α determines how much influence a single recommender has, and n is the number of recommenders. In this case, manipulation can be reduced, or even avoided in the briber setting, by scaling up the budget β proportionally to the competing incentive γ, reducing the influence α each recommender has, reducing the slope of the decision function ℓ, or increasing the number of recommenders. This last point is particularly interesting, as it suggests that one may reduce manipulation without increasing the total budget. Importantly, we did not assume that the principal has knowledge of what kind of information is available to recommenders or their prior beliefs. This makes our positive results widely applicable, and we hope they can guide the design of practical mechanisms for eliciting beliefs from experts with incentives that do not align with those of the principal.

REFERENCES

Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz. 2011. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*. 101–110.

B Douglas Bernheim. 1984. Rationalizable Strategic Behavior. *Econometrica: journal of the Econometric Society* 52, 4 (July 1984), 1007. https://www.jstor.org/stable/1911196?origin=crossref

Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

Yiling Chen, Ian Kash, Mike Ruberry, and Victor Shnayder. 2011. Decision Markets with Good Incentives. In *Internet and Network Economics*. Springer Berlin Heidelberg, 72–83.

Jacques Crémer and Richard P. McLean. 1988. Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions. *Econometrica* 56, 6 (1988), 1247–1257. http://www.jstor.org/stable/1913096

Edith Elkind, Piotr Faliszewski, and Arkadii Slinko. 2009. Swap bribery. In *International Symposium on Algorithmic Game Theory*. Springer, 299–310.

Piotr Faliszewski and Jörg Rothe. 2016. Control and Bribery in Voting.

Rupert Freeman and David M Pennock. 2018. An Axiomatic View of the Parimutuel Consensus Wagering Mechanism.. In *AAMAS*. 1936–1938.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

John Hopcroft and Daniel Sheldon. 2007. Manipulation-resistant reputations using hitting time. In *International Workshop on Algorithms and Models for the Web-Graph*. Springer, 68–81.

Leonard Hurwicz. 1960. Optimality and Informational Efficiency in Resource Allocation Processes. In *Mathematical Methods in the Social Sciences*. Stanford University Press, Stanford, CA, 27–46.

Philippe Jehiel, Moritz Meyer-ter Vehn, Benny Moldovanu, and William R Zame. 2006. The limits of ex post implementation. *Econometrica* 74, 3 (2006), 585–610.

Radu Jurca and Boi Faltings. 2009. Mechanisms for making crowds truthful. *J. Artif. Int. Res.* 34 (2009), 209–253.

Orgad Keller, Avinatan Hassidim, and Noam Hazon. 2018. Approximating bribery in scoring rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

David Kurokawa, Omer Lev, Jamie Morgenstern, and Ariel D Procaccia. 2015. Impartial peer review. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Claudio Mezzetti. 2004. Mechanism design with interdependent valuations: Efficiency. *Econometrica* 72, 5 (2004), 1617–1626.

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

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

David C Parkes and Sven Seuken. 2022. Economics and Computation - Book in Preparation. (2022).

David C. Parkes, Paul Tylkin, and Lirong Xia. 2017. Thwarting Vote Buying Through Decoy Ballots: Extended Version. In *Autonomous Agents and Multiagent Systems*. Springer International Publishing, Cham, 45–66.

David G Pearce. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica: journal of the Econometric Society* 52, 4 (July 1984), 1029. https://www.jstor.org/stable/1911197?origin=crossref

Drazen Prelec. 2004. A Bayesian truth serum for subjective data. *science* 306, 5695 (2004), 462–466.

Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

Robert L Winkler. 1994. Evaluating probabilities: Asymmetric scoring rules. *Management Science* 40, 11 (1994), 1395–1405.

Jens Witkowski and David C Parkes. 2012. A robust bayesian truth serum for small populations. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Justin Wolfers and Eric Zitzewitz. 2004. Prediction markets. *Journal of economic perspectives* 18, 2 (2004), 107–126.

Mark York, Munther Dahleh, and David C Parkes. 2021. Eliciting Social Knowledge for Creditworthiness Assessment. In *International Conference on Web and Internet Economics*. Springer, 428–445.

Haoqi Zhang, Eric Horvitz, Yiling Chen, and David C Parkes. 2011. Task routing for prediction tasks. (2011).

A CONSISTENT BELIEFS

In the briber setting, recommender i's subjective expected utility is (recall that s is a mixed strategy and we use $\hat{S}$ to denote the pure strategy samples from the mixed strategy):

$$
\begin{aligned}
&\textsc{eutil}_i\left(s, s^{\text{BRI}}, p_i, y_i, c_i\right) \\
&= \mathbb{E}\left[\textsc{util}_i\left(P_i, Y, C_i, R\right) \middle| s, s^{\text{BRI}}, p_i, y_i, c_i\right] \\
&= \mathbb{E}\left[\textsc{util}_i\left(p_i, \left(y_i, Y_{-i}\right), c_i, R\right) \middle| s, s^{\text{BRI}}, p_i, y_i, c_i\right] \\
&= \mathbb{E}\left[\textsc{util}_i\left(p_i, \left(y_i, Y_{-i}\right), c_i, \left(\hat{S}_i\left(p_i, y_i, c_i\right), \hat{S}_{-i}\left(P_{-i}, Y_{-i}, \hat{S}_{-i}^{\text{BRI}}(V, W)\right)\right)\right) \middle| s, s^{\text{BRI}}, p_i, y_i, c_i\right] \\
&= \mathbb{E}\left[\textsc{util}_i\left(p_i, \left(y_i, Y_{-i}\right), c_i, \left[\hat{S}_i\left(p_i, y_i, c_i\right), \hat{S}_{-i}\left(P_{-i}, Y_{-i}, \hat{S}_{-i}^{\text{BRI}}(V, W)\right)\right]\right) \middle| s, s^{\text{BRI}}, c_i\right]
\end{aligned}
$$

where the expectation is taken with respect to the conditional distribution that follows from Bayes' rule, i.e.

$$
\begin{aligned}
&\mathbb{P}\left[p_{-i}, y_{-i}, \hat{s}, \hat{s}_{-i}^{\text{BRI}}, v, w \middle| s, s^{\text{BRI}}, c_i\right] \\
&= \mathbb{P}\left[p_{-i}, y_{-i}, \hat{s}, \hat{s}_{-i}^{\text{BRI}} \middle| v, w, s, s^{\text{BRI}}, c_i\right] \mathbb{P}\left[\hat{s}_{-i}^{\text{BRI}}, v, w \middle| s, s^{\text{BRI}}, c_i\right] \\
\text{(independences)} &= \mathbb{P}\left[p_{-i}, y_{-i}, \hat{s}, \hat{s}_{-i}^{\text{BRI}} \middle| s, s^{\text{BRI}}\right] \mathbb{P}\left[v, w \middle| s^{\text{BRI}}, c_i\right].
\end{aligned}
$$

The only term that is potentially problematic is

$$
\mathbb{P}\left[v, w \middle| s, s^{\text{BRI}}, c_i\right].
$$

This distribution is not well-defined if there is no $\hat{s}^{\text{BRI}} \in \textsc{supp}(s^{\text{BRI}}), v \in \textsc{supp}(D^v), w \in \textsc{supp}(D^w)$ such that $c_i = \hat{s}_i^{\text{BRI}}(v, w)$, i.e. if we conditioned on a zero-probability event. In that case, any belief regarding the briber type, V, W, is said to be consistent and can hence be part of a perfect Bayesian equilibrium.

B SIMULATIONS

We here simulate the mechanisms from this paper to provide intuition for how they work, concrete budget requirements, and performance bounds for real-world deployments.

B.1 Simulation Setup

We base our parameter assumptions on an empirical lending study we conducted in collaboration with Makere University in Uganda and supported by the *Global Challenges in Economics and Computation Prize*. Through this partnership, we provided loans of approximately $80 USD to 75 smallholder farmers in Uganda, of whom 85% repaid in full. We thus assume a mean repayment probability of .85, and we assume that repayment is binary and drawn as a Bernoulli random trial. Since our data only include one draw per borrower, we have no estimate of borrower-level variance (the second moment). Fortunately, we gathered predictions of each borrower's repayment probability from 32 recommenders. We therefore estimate the distribution of borrower repayment probabilities by fitting a beta distribution to the borrower-average ratings from the study (see figure 1). Fitting the beta distribution with a maximum likelihood estimator, we find that $\alpha + \beta = 9.9$. We thus assume that each borrower i's Bernoulli repayment probability $\zeta_i = \mathbb{P}[O_i = 1]$ is identically

and independently distributed (i.i.d.) according to a beta distribution with $\alpha + \beta = 9.9$ and mean $\frac{\alpha}{\alpha+\beta} = .85$.

The mean of the recommenders' repayment probability predictions was .833, which is nearly centered on the mean repayment of .85. The variance of these predictions around the per-borrower means was .0101. For a beta distribution to have a mean of .833 and variance of .0101 would require $\alpha = 10.62$ and $\beta = 2.12$, meaning that $\alpha + \beta = 12.74$. We thus assume the same for the recommenders in our simulation. A more exact account of the experimental data would be to fit a maximum likelihood estimator, but this is non-standard since the parameters of the second beta distribution depend on the draw from the first. This would likely best be achieved with an iterative algorithm like Expectation Maximization, but the authors did not deem it necessary for the purposes of this simulation.

Formally, the generative model we implement in our simulation is:

$$\zeta_i \sim \mathcal{B}(\alpha, \beta), \quad \alpha = .85 \cdot 9.9, \beta = .15 \cdot 9.9 \tag{5}$$

$$Q_j \sim \mathcal{B}(\alpha_i, \beta_i), \quad \alpha_i = \zeta_i \cdot 12.74, \beta_i = \zeta_i \cdot (1 - 12.74) \tag{6}$$

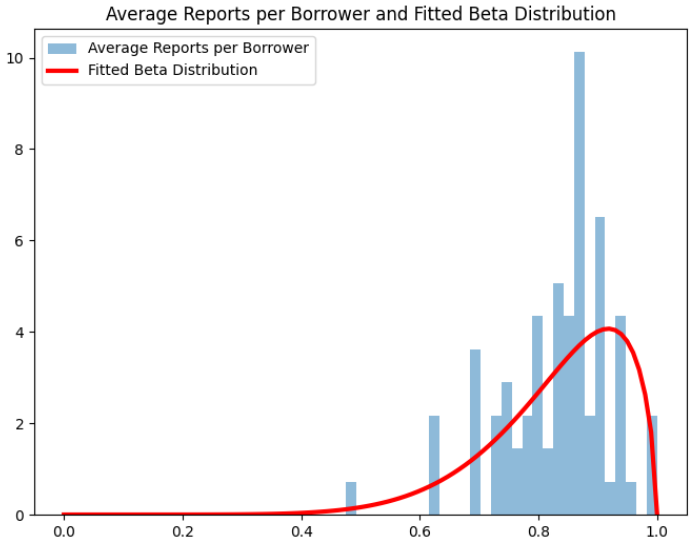

Fig. 1. Histogram of average per-borrower reports with fitted beta distribution

The principal elicits reports from n recommenders, make allocation decisions according to a bounded linear target decision function $f(R)$ as in definition 9. The principal pays each recommender j according to an adaptive payment mechanism $\text{PAY}(R, O)$ from definition 8. The principal receives a net utility u_i for borrower i of 1.2 times the allocation value in the case that $O = 1$ and 0 otherwise, less the sum of all payments made to recommenders.

$$u_i = 1.2 \cdot f(R) \cdot O - \sum_{j \in [n]} \text{PAY}(R, O) \tag{7}$$

We parameterize $f(R)$ such that $f(R) \in [0,1]$. We interpret $f(R)$ as the percentage of the requested loan that is given to the borrower, though we could equivalently interpret it as the probability that the borrower receives a loan of the full requested size. In our simulation, all recommenders' reports are influential (i.e. $m = n$). We introduce a parameter f_{min}, which is the minimum allocation that we give to each borrower no matter the reports q; this is part of the sufficient condition for truthfulness from Theorem 10. Finally, we set t as a parameter in the range $[0,1]$, $\bar{f} = \frac{1+f_{min}}{2}$, $\alpha = \varepsilon = \frac{1-f_{min}}{2}$, and

$$\ell = \begin{cases} \frac{\sqrt{n}}{.5} & n \leq 25 \\ 10 & n > 25 \end{cases} \tag{8}$$

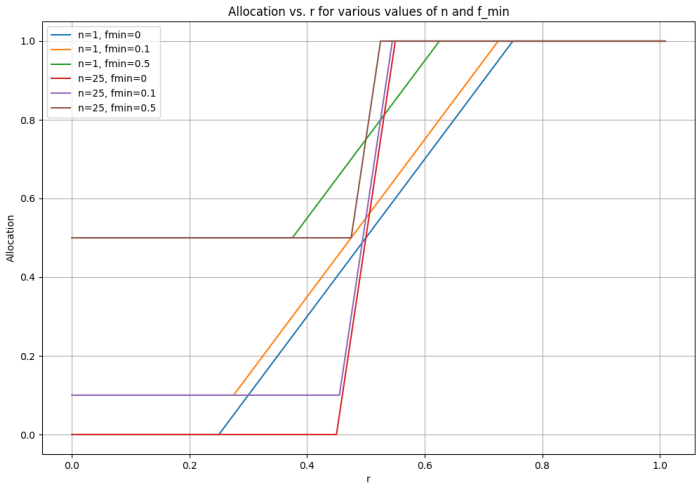

Fig. 2. Bounded Linear Target Allocation Function for various numbers of recommenders n, values f_{min}, and average reports r when all reports are equal.

This allocation function gives us an adaptive payment function of

$$\text{ADA-PAY}_i(r_i, o) = \left\{ \frac{\beta}{3n} \cdot \left(\frac{\int_0^{r_i} \delta_i(x)(o-x) \cdot dx}{\int_0^1 \delta_i(x) \cdot dx} + 1 \right) \right.$$

where $\delta_i(r_i)$ becomes

$$\delta_i(r_i) = \max_{r_{-i}} \left| \partial_{r_i} f(r) \right| + \int_0^1 \max_{r_{-i}} \left| \partial_{r_i} f(r) \right| dr_i = \frac{2\ell}{n}$$

resulting in

$$\text{ADA-PAY}_i(r_i, o) = \left\{ \frac{\beta}{3n} \cdot \left(\frac{\int_0^{r_i} \delta_i(x)(o-x) \cdot dx}{\int_0^1 \delta_i(x) \cdot dx} + 1 \right) \right. = \frac{\beta}{3n} \left(o r_i - \frac{r_i^2}{2} + 1 \right)$$

This is an asymmetric but strictly-proper scoring rule, assuming the outcome is always observed.

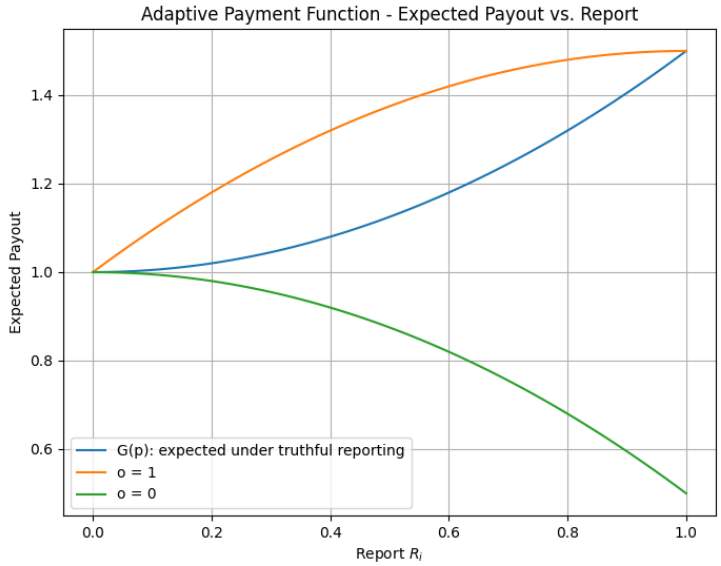

Fig. 3. Expected payout under the adaptive payment function under truthful reporting, and when $o = 1$ and $o = 0$. Note that $G(p)$ is strictly-convex, making the scoring rule strictly-proper.

B.2 Analytical Results

B.2.1 Required Budget β. We show analytically the budget required for various values of n and f_{min}, leveraging the sufficient condition in thm 10. We assume that the briber incentive $v = 1$, i.e. the briber's incentive is equal to the value of the loan. The charts in figure 4 and figure 5 show that the required budget β is multiple times the size of the loan unless we set a minimum allocation probability close to 1 or we have thousands of recommenders.

B.2.2 Accuracy. Given that the recommenders' beliefs q_i are independently distributed with beta distribution variance of $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, we have that

$$Var\left[\frac{1}{n}\sum_{j}^{n} q_j\right] = \frac{Var[q_j]}{n} \tag{9}$$

$$= \frac{\alpha\beta}{n(\alpha+\beta)^2(\alpha+\beta+1)} \tag{10}$$

Given that $\alpha = 18\zeta_i$ and $\beta = 18(1-\zeta_i)$, the value of this variance of an estimate based on the average of all n true beliefs q_i is

$$\int_0^1 \mathbb{P}[\zeta_i] \frac{\alpha\beta}{n(\alpha+\beta)^2(\alpha+\beta+1)} d\zeta_i = \int_0^1 \mathbb{P}[\zeta_i] \frac{\zeta_i(1-\zeta_i)}{19n} d\zeta_i \tag{11}$$

If we assume that all borrowers have the same repayment probability of .85, the above integral becomes $\sim \frac{.0067}{n}$, making the standard deviation $\sim \frac{.08919}{\sqrt{n}}$. In figure 6, we chart this line versus

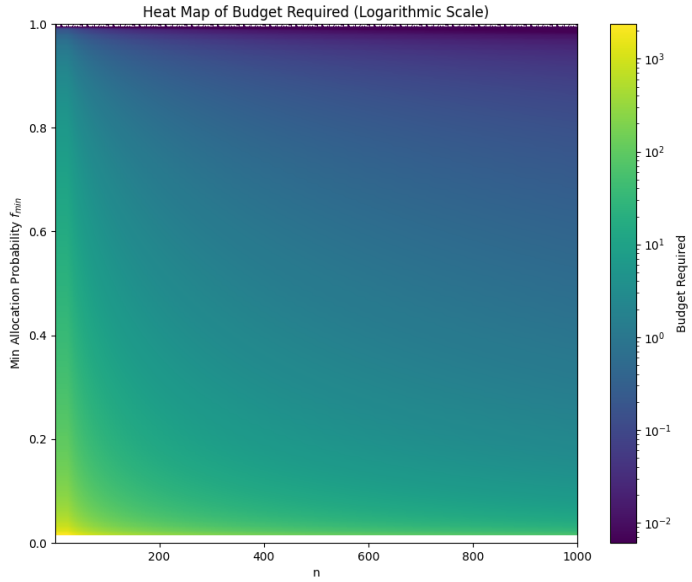

Fig. 4. Required budget β for various numbers of recommenders n and minimum allocations f_{min}. Assuming that the briber incentive v is equivalent to the loan value, the lender will require a budget larger than the loan value unless the number of recomemenders n is in the thousands.

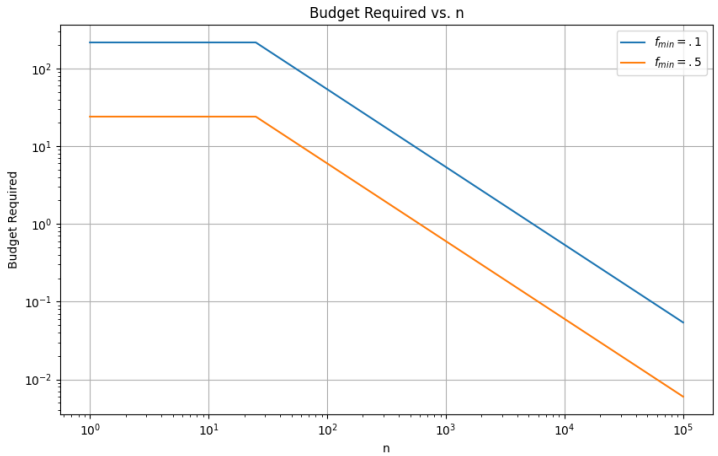

Fig. 5. Required budget β for various numbers of recommenders n and $f_{min} \in \{.1, .5\}$.

the empirically-observed error for n agents in 10,000 runs in simulation. We see that just ten recommenders already give reasonable accuracy.

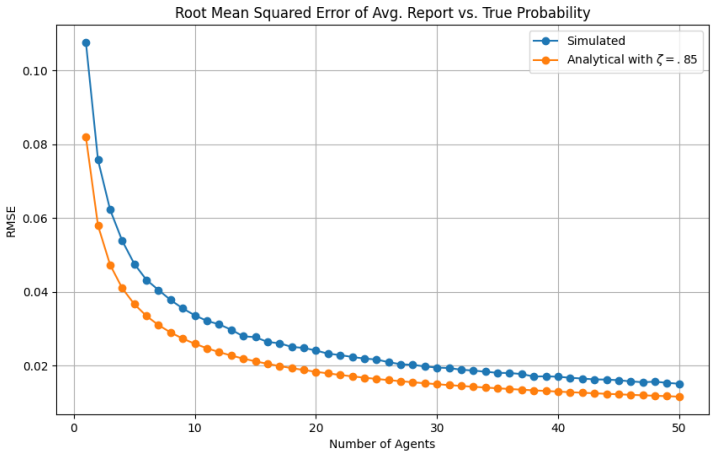

Fig. 6. Accuracy for repayment probability when averaging truthful reports from recommenders. Simulated values include a range of true repayment probabilities, vs. the analytical value where repayment probability is always .85.

B.3 Simulated Results

From (11), it is clear that accuracy increases with n. Likewise, per the sufficient condition in thm. 10, the required budget β decreases with n. Therefore the lender will prefer n to be as large as possible, and we focus our analysis on the impact of changing the minimum allocation probability f_{min} and the center of the allocation function t to find the profit-optimal setting.

In the heatmap in figure 7, we simulated a setting with $n = 100,000$ recommenders and an averaged across 10,000 runs for each combination of t and f_{min} values. We see empirically that a setting of $t = .84$ and $f_{min} = .32$ is profit optimal, giving a profit of 4.5% of loan value.

In figure 8, we run the same simulation for 10^6 recommenders. We observe that the profit-optimal settings are $t = .82$ and $f_{min} = .12$, while the ROI-optimal $t = .92$ and $f_{min} = .06$. This makes sense as lending to lower-quality borrowers can still yield a small profit, but the expected return on capital deployed is lower. Note that we were not able to obtain a profitable setting with fewer than 1,000 recommenders, and that all profitable settings with fewer than 100,000 recommenders effectively ignore recommender reports (i.e. $f_{min} \sim 1$).

In figure 9, we see that both profit and ROI are monotonically-increasing in n under optimal f_{min} and t for at least $n \in [1, 10^6]$.

B.4 Discussion

We showed analytically that accuracy and required budget improve with increasing n. Our simulation showed that a minimum of around 100,000 recommenders is required for the lender to profitably use recommenders (given our starting assumptions), and that there are optimal values of f_{min} and t which must be found before deploying such a lending system.

From a practical standpoint, it would be difficult in many settings to source 100,000 recommenders. The most likely settings where this could work are large-scale prediction tasks and prediction

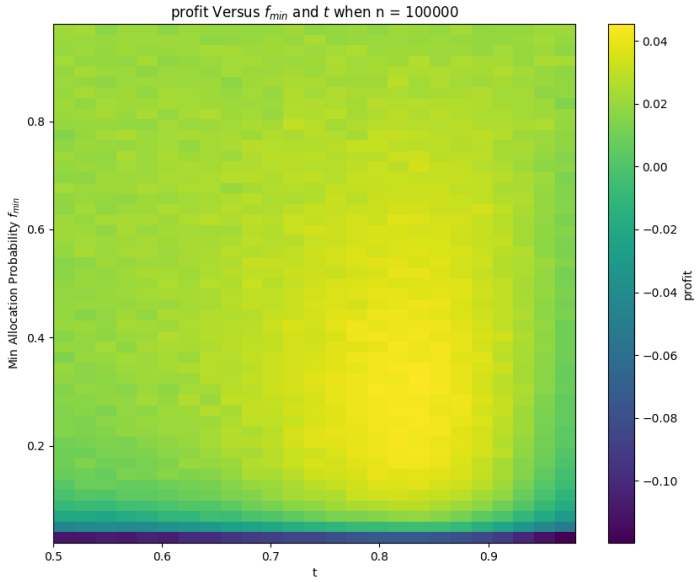

Fig. 7. Expected profit with 100,000 recommenders for various values of f_{min} and t. Grid cell values are the average of 10,000 simulation runs.

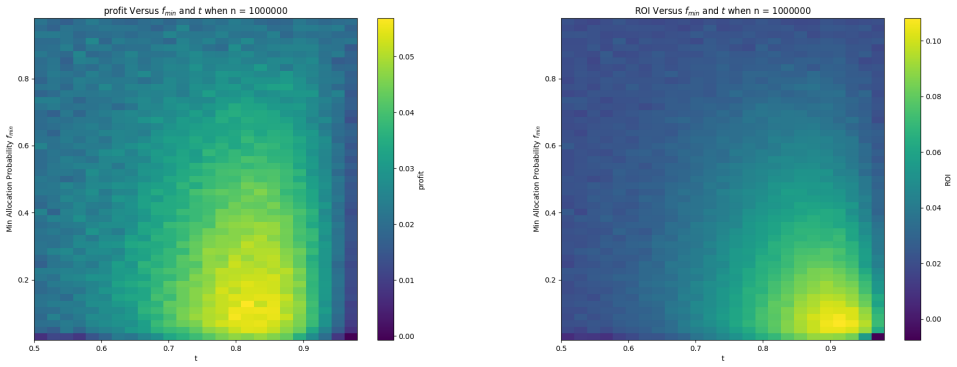

Fig. 8. Expected profit and ROI with 1,000,000 recommenders for various values of f_{min} and t. Grid cell values are the average of 10,000 simulation runs. We see that the max ROI setting has a higher t and lower f_{min} value than the max total profit setting.

markets. It could also work if the briber's incentive is much less than the amount at stake for the principal; in our case the amounts are equal. Another consideration is that the expected score for a recommender would be so minimal with a smaller total budget divided by a large n that it may not be salient enough to incentivize effort (per the literature on incentivizing agents to invest effort to

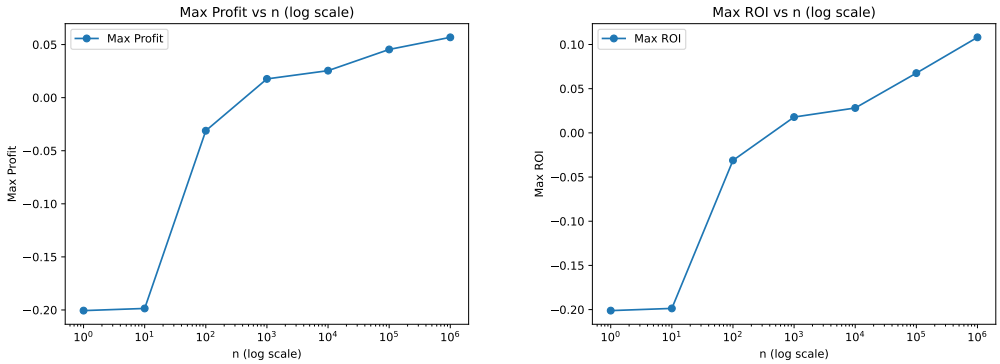

Fig. 9. Expected profit and ROI under the empirically-optimal settings of f_{min} and t for various values of n. The second increase in slope happens when it becomes economically-viable for the mechanism to use recommender reports with a low f_{min} around $n = 10^5$

obtain a high-quality signal). Nonetheless with sufficent n and a well-tuned allocation function, this system could find use in some settings.

C PROOFS

C.1 Proof of Theorem 6

We start by proving a few results that we will need to prove Theorem 6.

Lemma 5. *Suppose there is a list of functions, $f_1, ..., f_n : [0,1]^n \to \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^n f_i(x) \leq \beta \; \forall x \in \mathbb{R}^n$, and $\mathcal{W} = (\mathcal{K}, \mathcal{D}_0, \mathcal{D}_z)$ is some walk Definition 5. Then, with $k = |\mathcal{K}|$, we have*

$$\sum_{m=1}^{k} \mathbb{E}_{X\sim\mathcal{W}} \left[f_{[X^m \neq X^{m-1}]} \left(X^m\right) + f_{[X^m \neq X^{m-1}]} \left(X^{m-1}\right) \right] \leq 2 \cdot \beta \cdot H(k)$$

where H is the harmonic number.

Proof. Dropping $X \sim \mathcal{W}$, and using the random variables as defined in Definition 5, we have

$$\mathbb{E}\left[f_{[X^m \neq X^{m-1}]} \left(X^m\right) + f_{[X^m \neq X^{m-1}]} \left(X^{m-1}\right) \right] = \mathbb{E}\left[f_{\Theta^m} \left(X^m\right) + f_{\Theta^m} \left(X^{m-1}\right) \right]$$

For convenience, let's define a few functions: For any set $T \subseteq \{1, .., k\}$, let's define the random variables

$$Y_i(T) := \begin{cases} X_i^0 + Z_i & \text{if } i \in T \\ X_i^0 & \text{else.} \end{cases}$$

Further, Σ be the set of all permutations on $\mathcal{K}$ and let for any $\sigma \in \Sigma$

$$S^\sigma(m) := \{\sigma(1), .., \sigma(m)\}.$$

We hence have

$$X^j = Y\left(S^\Theta(j)\right).$$

Defining

$$\begin{aligned} g_{\Theta^m}\left(S^\Theta(j)\right) :&= f_{\Theta^m}\left(Y\left(S^\Theta(j)\right)\right) \\ &= f_{\Theta^m}\left(X^j\right) \end{aligned}$$

we have

$$\mathbb{E}\left[f_{\Theta^m}\left(X^m\right) + f_{\Theta^m}\left(X^{m-1}\right)\right] = \mathbb{E}\left[g_{\Theta^m}\left(S^\Theta(m)\right)\right] + \mathbb{E}\left[g_{\Theta^m}\left(S^\Theta(m-1)\right)\right].$$

We will now prove that

$$\sum_{m=1}^{k} \mathbb{E}\left[g_{\Theta^m}\left(S^\Theta(m-1)\right)\middle| x^0, z\right] \leq H(k)\cdot\beta \quad \forall x^0, z$$

which implies that

$$\sum_{m=1}^{k} \mathbb{E}\left[g_{\Theta^m}\left(S^\Theta(m-1)\right)\right] \leq H(k)\cdot\beta.$$

Note that $\mathbb{I}[\cdot]$ is the indicator function. We have

$$\begin{aligned}
&\sum_{m=1}^{k} \mathbb{E}\left[g_{\Theta^m}\left(S^\Theta(m-1)\right)\middle| x^0, z\right]\cdot k! \\
&= \sum_{\sigma\in\Sigma}\sum_{m=1}^{k} g_{\sigma(m)}\left(S^\sigma(m-1)\right) \\
&= \sum_{\sigma\in\Sigma}\sum_{m=1}^{k}\sum_{n=1}^{k}\sum_{Z}\mathbb{I}\left[\sigma(m)=n \text{ and } Z=S^\sigma(m-1)\right] g_n(Z) \\
&= \sum_{n=1}^{k}\sum_{Z} g_n(Z)\sum_{m=1}^{k}\sum_{\sigma\in\Sigma}\mathbb{I}\left[\sigma(m)=n \text{ and } Z=S^\sigma(m-1)\right] \\
&= \sum_{n=1}^{k}\sum_{Z} g_n(Z)\sum_{m=1}^{k}\sum_{\sigma\in\Sigma}\mathbb{I}\left[\sigma(m)=n \text{ and } Z=S^\sigma(m-1)\right]\mathbb{I}\left[|Z|=m-1\right]\mathbb{I}\left[n\notin Z\right].
\end{aligned}$$

Note that the first indicator function requires that the first $m-1$ elements of the permutation σ are Z, the m-th element is n, and the last $k-m$ elements are $(\{1,..,k\}\setminus Z)\setminus\{n\}$. The degrees of freedom of such a permutation are the order of the first $m-1$ elements and the order of the last $k-m$ elements, we hence have

$$\begin{aligned}
&= \sum_{n=1}^{k}\sum_{Z} g_n(Z)\sum_{m=1}^{k}(m-1)!(k-m)!\mathbb{I}\left[|Z|=m-1\right]\mathbb{I}\left[n\notin Z\right] \\
&\leq \sum_{Z}\sum_{m=1}^{k}(m-1)!(k-m)!\mathbb{I}\left[|Z|=m-1\right]\sum_{n=1}^{k} g_n(Z)
\end{aligned}$$

Since $\sum_{i=1}^{n} f_i(x) \leq \beta$, we have $\sum_{n=1}^{k} g_n(Z) \leq \beta$ and hence

$$\begin{aligned}
\sum_{\sigma\in\Sigma}\sum_{m=1}^{k} g_{\sigma(m)}\left(S^{\sigma}(m-1)\right) &\leq \sum_{Z}\sum_{m=1}^{k}(m-1)!(k-m)!\mathbb{I}\left[|Z|=m-1\right]\cdot\beta \\
&= \sum_{m=1}^{k}(m-1)!(k-m)!\binom{k}{m-1}\cdot\beta \\
&= \sum_{m=1}^{k}\frac{k!}{(k-m+1)}\cdot\beta \\
&= \sum_{m=1}^{k}\frac{k!}{m}\cdot\beta \\
&= k!\cdot H(k)\cdot\beta.
\end{aligned}$$

For the other term, we can make a similar argument:

$$\begin{aligned}
&\sum_{m=1}^{k}\mathbb{E}\left[g_{\Theta^m}\left(S^{\Theta}(m)\right)\middle|x^0,z\right]\cdot k! \\
&= \sum_{\sigma\in\Sigma}\sum_{m=1}^{k} g_{\sigma(m)}\left(S^{\sigma}(m)\right) \\
&= \sum_{\sigma\in\Sigma}\sum_{m=1}^{k}\sum_{n=1}^{k}\sum_{Z}\mathbb{I}\left[\sigma(m)=n \text{ and } Z=S^{\sigma}(m)\right] g_n(Z) \\
&= \sum_{n=1}^{k}\sum_{Z} g_n(Z)\sum_{m=1}^{k}\sum_{\sigma\in\Sigma}\mathbb{I}\left[\sigma(m)=n \text{ and } Z=S^{\sigma}(m)\right] \\
&= \sum_{n=1}^{k}\sum_{Z} g_n(Z)\sum_{m=1}^{k}\sum_{\sigma\in\Sigma}\mathbb{I}\left[\sigma(m)=n \text{ and } Z=S^{\sigma}(m)\right]\mathbb{I}\left[|Z|=m\right]\mathbb{I}\left[n\notin Z\right] \\
&= \sum_{n=1}^{k}\sum_{Z} g_n(Z)\sum_{m=1}^{k}(m-1)!(k-m)!\mathbb{I}\left[|Z|=m\right]\mathbb{I}\left[n\notin Z\right] \\
&\leq \sum_{Z}\sum_{m=1}^{k}(m-1)!(k-m)!\mathbb{I}\left[|Z|=m\right]\sum_{n=1}^{k} g_n(Z) \\
&\leq \sum_{Z}\sum_{m=1}^{k}(m-1)!(k-m)!\mathbb{I}\left[|Z|=m\right]\cdot\beta \\
&= \sum_{m=1}^{k}(m-1)!(k-m)!\binom{k}{m}\cdot\beta \\
&= \sum_{m=1}^{k}\frac{k!}{m}\cdot\beta \\
&= k!\cdot H(k)\cdot\beta.
\end{aligned}$$

The result follows straightforwardly. □

Lemma 6. *For any proper scoring rule* SC*, with* $\beta \geq \mathrm{SC}(x,y) \geq 0$, $\forall x,y \in [0,1]$*, and any monotone sequence* $x^{(0)},..,x^{(K)} \in [0,1]$*, such that* $x^{(k+1)} - x^{(k)} = x^{(j+1)} - x^{(j)}$ $\forall k,j \in [K-1]$*, we have*

$$\sum_{k=1}^{K} \text{D-SC}\left(x^{(k-1)}, x^{(k)}\right) = \frac{1}{K} \cdot \text{D-SC}(x^{(0)}, x^{(K)})$$

with

$$\text{D-SC}(a,b) := \mathrm{SC}(a,a) - \mathrm{SC}(b,a) + \mathrm{SC}(b,b) - \mathrm{SC}(a,b).$$

Proof. Any scoring rule for dichotomous RVs can be written as

$$\mathrm{sc}(x,r) = g(x)r + h(x)(1-r),$$

for some $g(x)$ and $h(x)$. Using the identity

$$\begin{aligned}\mathrm{sc}(a,b) - \mathrm{sc}(a,c) &= g(a)b + h(a)(1-b) - g(a)c - h(a)(1-c)\\ &= (g(a) - h(a))\,(b-c),\end{aligned}$$

we have

$$\begin{aligned}&\mathrm{sc}(x,r) - \mathrm{sc}(y,r)\\ &= (\mathrm{sc}(x,r) - \mathrm{sc}(x,x)) + (\mathrm{sc}(y,x) - \mathrm{sc}(y,r)) + \mathrm{sc}(x,x) - \mathrm{sc}(y,x)\\ &= (g(x) - h(x))\,(r-x) + (g(y) - h(y))\,(x-r) + \mathrm{sc}(x,x) - \mathrm{sc}(y,x)\\ &= ((g(x) - h(x))\,(x-y) + (g(y) - h(y))\,(y-x))\,\frac{x-r}{y-x} + \mathrm{sc}(x,x) - \mathrm{sc}(y,x)\\ &= (\mathrm{sc}(x,x) - \mathrm{sc}(x,y) + \mathrm{sc}(y,y) - \mathrm{sc}(y,x))\,\frac{x-r}{y-x} + \mathrm{sc}(x,x) - \mathrm{sc}(y,x)\\ &= (\mathrm{sc}(x,x) - \mathrm{sc}(y,x))\,\frac{y-r}{y-x} + (\mathrm{sc}(y,y) - \mathrm{sc}(x,y))\,\frac{x-r}{y-x}.\end{aligned}$$

With $r = 0$, we have

$$\mathrm{sc}(x,0) - \mathrm{sc}(y,0) = (\mathrm{sc}(x,x) - \mathrm{sc}(y,x))\,\frac{y}{y-x} + (\mathrm{sc}(y,y) - \mathrm{sc}(x,y))\,\frac{x}{y-x},$$

and with $r = 1$, we have

$$\mathrm{sc}(x,1) - \mathrm{sc}(y,1) = (\mathrm{sc}(x,x) - \mathrm{sc}(y,x))\,\frac{y-1}{y-x} + (\mathrm{sc}(y,y) - \mathrm{sc}(x,y))\,\frac{x-1}{y-x}.$$

Subtracting the second equality from the first, we obtain

$$\begin{aligned}&\frac{\mathrm{sc}(x,x) - \mathrm{sc}(y,x) + \mathrm{sc}(y,y) - \mathrm{sc}(x,y)}{y-x} = \mathrm{sc}(x,0) - \mathrm{sc}(y,0) - (\mathrm{sc}(x,1) - \mathrm{sc}(y,1))\\ &\mathrm{sc}(x,x) - \mathrm{sc}(y,x) + \mathrm{sc}(y,y) - \mathrm{sc}(x,y) = (y-x)\,(\mathrm{sc}(x,0) - \mathrm{sc}(y,0) - \mathrm{sc}(x,1) + \mathrm{sc}(y,1))\,. \quad (12)\end{aligned}$$

Now, for an increasing sequence $x^{(0)}, .., x^{(K)}$ with $|x^{(k+1)} - x^{(k)}| = \frac{|x^{(K)} - x^{(0)}|}{K} \quad \forall k$, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \text{D-SC}(x^{(k)}, x^{(k+1)}) \\
&= \sum_{k=0}^{K-1} \left(\text{SC}(x^{(k)}, x^{(k)}) - \text{SC}(x^{(k+1)}, x^{(k)}) + \text{SC}(x^{(k+1)}, x^{(k+1)}) - \text{SC}(x^{(k)}, x^{(k+1)}) \right) \\
&= \sum_{k=0}^{K-1} (x^{(k+1)} - x^{(k)}) \overbrace{\left(\text{SC}(x^{(k)}, 0) - \text{SC}(x^{(k+1)}, 0) - \text{SC}(x^{(k)}, 1) + \text{SC}(x^{(k+1)}, 1) \right)}^{\geq 0} \text{ (From(12))} \\
&= \frac{\left|x^{(K)} - x^{(0)}\right|}{K} \cdot \sum_{k=1}^{K-1} \left(\text{SC}(x^{(k)}, 0) - \text{SC}(x^{(k+1)}, 0) - \text{SC}(x^{(k)}, 1) + \text{SC}(x^{(k+1)}, 1) \right) \\
&= \frac{\left|x^{(K)} - x^{(0)}\right|}{K} \cdot \left(\text{SC}(x^{(0)}, 0) - \text{SC}(x^{(K)}, 0) - \text{SC}(x^{(0)}, 1) + \text{SC}(x^{(K)}, 1) \right) \\
&= \frac{1}{K} \cdot \left(\text{SC}(x^{(0)}, x^{(0)}) - \text{SC}(x^{(K)}, x^{(0)}) - \text{SC}(x^{(0)}, x^{(K)}) + \text{SC}(x^{(K)}, x^{(K)}) \right) \\
&= \frac{1}{K} \cdot \text{D-SC}(x^{(0)}, x^{(K)})
\end{aligned}$$

and similarly for a decreasing sequence. Hence, the result follows. □

Lemma 7. *For any mechanism* $\text{MECH} = (F, \text{PAY})$ *with budget* $\beta \geq 0$, *any* $\gamma \geq 0$, $k \in [n]$,$\delta \in [0, 1]$, *there exists a* $q \in \mathcal{Q}$, $t \in [n]$, $x \in [0, 1]$, $c_t \in \{-\gamma, \gamma\}$, *with* $|q_t - x| \leq \delta$, *such that*

$$\frac{\gamma \cdot \Lambda_F(k, \delta)^2}{8 \cdot k \cdot \delta \cdot \frac{\beta}{\gamma} \cdot H(k) + 4 \cdot k \cdot \Lambda_F(k, \delta)} \leq \mathbb{E}_{O \sim q_t} \left[\text{PAY}_t \left((x, q_{-t}), O \right) - \text{PAY}_t(q, O) \right] + c_t \left(F(x, q_{-t}) - F(q) \right)$$

where H *is the harmonic number.*

Proof. For convenience, we will sometimes use the notation

$$\begin{aligned}
\text{PAY}_i(r_i, q_i | r_{-i}) &:= \mathbb{E}_{O \sim q_i} \left[\text{PAY}_i(r, O) \right] \\
&= \underbrace{\text{PAY}_i(r, 1)}_{=: g_i(r)} \cdot q_i + \underbrace{\text{PAY}_i(r, 0)}_{=: h_i(r)} \cdot (1 - q_i).
\end{aligned}$$

Note that $\text{PAY}_i(r_i, o | r_{-i}) = \text{PAY}_i(r, o) \ \forall o \in \mathcal{O}$. By definition of our mechanism (Definition 1), $\text{PAY}_i(r_i, q_i | r_{-i})$ is a strictly proper scoring rule in r_i, q_i. For any $a, b \in \mathcal{R}$ that differ only at index i, let

$$\text{D-PAY}(a, b) := \text{PAY}_i(a_i, a_i | a_{-i}) - \text{PAY}_i(b_i, a_i | a_{-i}) + \text{PAY}_i(b_i, b_i | a_{-i}) - \text{PAY}_i(a_i, b_i | a_{-i})$$

which we can interpret as a notion of the cost of misreporting, we shall see below how we use this quantity. We have

$$\begin{aligned}
&\text{D-PAY}(a, b) \\
&= g(a) a_i + h(a)(1 - a_i) - g(b) a_i - h(b)(1 - a_i) + g(b) b_i + h(b)(1 - b_i) - g(a) b_i - h(a)(1 - b_i) \\
&= \left(g(a) - h(a) - g(b) + h(b) \right) a_i + h(a) - h(b) + \left(g(b) - h(b) - g(a) + h(a) \right) b_i + h(b) - h(a) \\
&= \left(g(a) - h(a) - g(b) + h(b) \right) a_i + \left(g(b) - h(b) - g(a) + h(a) \right) b_i \\
&= \left(b_i - a_i \right) \left(g_i(b) - g_i(a) + h_i(a) - h_i(b) \right).
\end{aligned}$$

For any $\mathcal{W} \in \textsc{walks}(k, \delta)$, we have

$$\begin{aligned}
\textsc{cost} :&= \mathbb{E}_{x\sim\mathcal{W}}\left[\sum_{m=1}^{k} \textsc{d-pay}(x^{m-1}, x^m)\right] \\
&= \sum_{m=1}^{k} \mathbb{E}_{x\sim\mathcal{W}}\left[\left\|x^m - x^{m-1}\right\|_1 \left(g_{[x^m\neq x^{m-1}]}\left(x^m\right) - g_{[x^m\neq x^{m-1}]}\left(x^{m-1}\right)\right)\right] \\
&+ \sum_{m=1}^{k} \mathbb{E}_{x\sim\mathcal{W}}\left[\left\|x^m - x^{m-1}\right\|_1 \left(h_{[x^m\neq x^{m-1}]}\left(x^{m-1}\right) - h_{[x^m\neq x^{m-1}]}\left(x^m\right)\right)\right] \\
&\leq \delta \cdot \sum_{m=1}^{k} \mathbb{E}_{x\sim\mathcal{W}}\left[g_{[x^m\neq x^{m-1}]}\left(x^m\right) - g_{[x^m\neq x^{m-1}]}\left(x^{m-1}\right) + h_{[x^m\neq x^{m-1}]}\left(x^{m-1}\right) - h_{[x^m\neq x^{m-1}]}\left(x^m\right)\right] \\
&\leq \delta \cdot \sum_{m=1}^{k} \mathbb{E}_{x\sim\mathcal{W}}\left[g_{[x^m\neq x^{m-1}]}\left(x^m\right) + g_{[x^m\neq x^{m-1}]}\left(x^{m-1}\right)\right] \\
&+ \delta \cdot \sum_{m=1}^{k} \mathbb{E}_{x\sim\mathcal{W}}\left[h_{[x^m\neq x^{m-1}]}\left(x^{m-1}\right) + h_{[x^m\neq x^{m-1}]}\left(x^m\right)\right]
\end{aligned}$$

Since $g_i(x) = \textsc{pay}_i(x, 1)$ we have

$$\sum_{i=1}^{k} g_i(x) \leq \beta \quad \forall x$$

and similarly for h, hence we can apply Lemma 5 twice and we have

$$\textsc{cost} \leq 4 \cdot \delta \cdot \beta \cdot H(k).$$

Let $\Lambda(k, \delta)$ be the sensitivity function from Definition 6 and $\mathcal{W}$ its walk. We will sometimes simply write Λ without its arguments, for simplicity. We hence have for any $\omega \geq 0$

$$\begin{aligned}
\omega \cdot \Lambda - 2 \cdot \delta \cdot \beta \cdot H(k) &\leq \omega \cdot \mathbb{E}_{X\sim\mathcal{W}}\left[\sum_{i=1}^{k} \left|\mathrm{F}(x^i) - \mathrm{F}(x^{i-1})\right|\right] - \frac{1}{2}\textsc{cost} \\
&= \omega \cdot \mathbb{E}_{X\sim\mathcal{W}}\left[\sum_{i=1}^{k} \left|\mathrm{F}(x^i) - \mathrm{F}(x^{i-1})\right|\right] - \frac{1}{2}\mathbb{E}_{x\sim\mathcal{W}}\left[\sum_{m=1}^{k} \textsc{d-pay}(x^{m-1}, x^m)\right]
\end{aligned}$$

By the pidgeonhole principle, there must be a path x and a step m such that

$$\omega \cdot \Lambda - 2 \cdot \delta \cdot \beta \cdot H(k) \leq k\left(\omega\left|\mathrm{F}(x^m) - \mathrm{F}(x^{m-1})\right| - \frac{1}{2}\textsc{d-pay}(x^{m-1}, x^m)\right).$$

Let now $z^0, ..., z^J$ be the linear interpolation between x^{m-1} and x^m. We have

$$\begin{aligned}
\textsc{d-pay}(x^{m-1}, x^m) &= \textsc{d-pay}(z^0, z^J) \\
&= J \cdot \sum_{j=1}^{J} \textsc{d-pay}(z^{j-1}, z^j)
\end{aligned}$$

where the last equality follows from Lemma 6. Plugging this in, we have

$$\frac{1}{k}\left(\Lambda\cdot\omega-2\cdot\delta\cdot\beta\cdot H(k)\right)\leq\omega\left|\sum_{j=0}^{J-1}\left(\mathrm{F}\left(z^{j+1}\right)-\mathrm{F}\left(z^{j}\right)\right)\right|-\frac{1}{2}J\cdot\sum_{j=1}^{J}\text{D-PAY}(z^{j-1},z^{j})$$

$$\frac{1}{kJ^2}\left(\Lambda\cdot\omega-2\cdot\delta\cdot\beta\cdot H(k)\right)\leq\frac{1}{J}\sum_{j=0}^{J-1}\left(\frac{\omega}{J}\left|\left(\mathrm{F}\left(z^{j+1}\right)-\mathrm{F}\left(z^{j}\right)\right)\right|-\frac{1}{2}\text{D-PAY}(z^{j},z^{j+1}\right).$$

Applying the pidgeonhole principle again, it follows that there are report profiles z, z' such that (i) they differ only in one dimension (by at most δ) and (ii)

$$\frac{1}{kJ^2}\left(\Lambda\cdot\omega-2\cdot\delta\cdot\beta\cdot H(k)\right)\leq\frac{\omega}{J}\cdot\operatorname{sign}\left(\mathrm{F}\left(z'\right)-\mathrm{F}\left(z\right)\right)\cdot\left(\mathrm{F}\left(z'\right)-\mathrm{F}\left(z\right)\right)-\frac{1}{2}\text{D-PAY}(z,z').$$

With $c:=\frac{\omega}{J}\cdot\operatorname{sign}\left(\mathrm{F}\left(z'\right)-\mathrm{F}\left(z\right)\right)$, we have

$$\frac{1}{kJ^2}\left(\Lambda\cdot|c|\cdot J-2\cdot\delta\cdot\beta\cdot H(k)\right)\leq c\cdot\left(\mathrm{F}\left(z'\right)-\mathrm{F}\left(z\right)\right)-\frac{1}{2}\text{D-PAY}(z,z').$$

$$\frac{1}{kJ}\left(\Lambda\cdot|c|-\frac{1}{J}\cdot2\cdot\delta\cdot\beta\cdot H(k)\right)\leq c\left(\mathrm{F}\left(z'\right)-\mathrm{F}\left(z\right)\right)-\frac{1}{2}\text{D-PAY}(z,z')$$

This holds for any $J\in\mathbb{N}_{>0}$, let's pick

$$\begin{aligned}J&=\left\lceil\frac{4\cdot\delta\cdot\beta\cdot H(k)}{c\cdot\Lambda}\right\rceil\\&=\frac{4\cdot\delta\cdot\beta\cdot H(k)}{c\cdot\Lambda}+\underbrace{\left\lceil\frac{4\cdot\delta\cdot\beta\cdot H(k)}{c\cdot\Lambda}\right\rceil-\frac{4\cdot\delta\cdot\beta\cdot H(k)}{c\cdot\Lambda}}_{=:\epsilon}\end{aligned}$$

where $\lceil\cdot\rceil$ means rounding up to the next integer. With this J, we have

$$\begin{aligned}\frac{1}{kJ}\left(\Lambda-\frac{1}{J\cdot c}\cdot2\cdot\delta\cdot\beta\cdot H(k)\right)&=\frac{c\cdot\Lambda^2\cdot\left(2\cdot\delta\cdot\beta\cdot H(k)+\epsilon\cdot c\cdot\Lambda\right)}{k\cdot\left(4\cdot\delta\cdot\beta\cdot H(k)+\epsilon\cdot c\cdot\Lambda\right)^2}\\&\geq\frac{c\cdot\Lambda^2\cdot\left(2\cdot\delta\cdot\beta\cdot H(k)+\epsilon\cdot c\cdot\Lambda\right)}{k\cdot\left(4\cdot\delta\cdot\beta\cdot H(k)+2\cdot\epsilon\cdot c\cdot\Lambda\right)^2}\\&=\frac{c\cdot\Lambda^2}{k\cdot4\cdot\left(2\cdot\delta\cdot\beta\cdot H(k)+\epsilon\cdot c\cdot\Lambda\right)}\\&\geq\frac{c\cdot\Lambda^2}{k\cdot4\cdot\left(2\cdot\delta\cdot\beta\cdot H(k)+c\cdot\Lambda\right)}\end{aligned}$$

Hence, for any $\gamma\geq0$, there is a q, an i, a $c_i\in-\gamma,\gamma$ and an r_i such that

$$\begin{aligned}\frac{\gamma\cdot\Lambda^2}{k\cdot4\cdot\left(2\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+\Lambda\right)}&\leq-\frac{1}{2}\text{D-PAY}((q_{-i},r_i),q)+c_i\left(\mathrm{F}\left((q_{-i},r_i)\right)-\mathrm{F}\left(q\right)\right)\\&\leq\text{PAY}_i(r_i,q_i|q_{\neg i})-\text{PAY}_i(q_i,q_i|q_{\neg i})+c_i\left(\mathrm{F}\left((q_{-i},r_i)\right)-\mathrm{F}\left(q\right)\right).\end{aligned}$$

□

We are now ready to prove Theorem 6:

Proof. Fix some $k \in [n]$,$\delta \in [0,1]$, and let q, t, x, c_t as in Lemma 7. Suppose that the prior D only has support at

$$P_i(o|y_i) = p_i(o|y_i) := \begin{cases} q_i & \text{if } o = 1 \\ 1 - q_i & \text{if } o = 0 \end{cases} \quad \forall y_i$$
$$C_i = c_i := \begin{cases} c_t & i = t \\ 0 & i \neq t \end{cases}$$
$$Y = y := 0.$$

Then

$$\begin{aligned} \textsc{util}_i\left(p, y, c_i, r\right) &= \mathbb{E}_{O \sim p_i(\cdot|y)}\left[\textsc{pay}_i(r, O) + c_i \cdot \mathrm{F}(r)\right] \\ &= \textsc{pay}_i(r_i, q_i | r_{-i}) + c_i \cdot \mathrm{F}(r). \end{aligned}$$

Hence, we have

$$\textsc{eutil}_i\left(s, p_i, y_i, c_i\right) := \mathbb{E}\left[\textsc{util}_i\left(P_i, Y, C_i, R\right) \middle| s, p_i, y_i, c_i\right]$$

Note that since all priors have only point-support, this is not an incomplete-information game anymore, and we can identify strategies with distributions over reports, we hence write, with slight abuse of notation,

$$\textsc{eutil}_i\left(s, p_i, y_i, c_i\right) = \mathbb{E}_{R \sim s}\left[\textsc{pay}_i\left(R_i, q_i \middle| R_{-i}\right) + c_i \cdot \mathrm{F}\left(R\right)\right]. \tag{13}$$

For any $i \neq t$ we have

$$\textsc{eutil}_i\left(s, p_i, y_i, c_i\right) = \mathbb{E}_{R \sim s}\left[\textsc{pay}_i\left(R_i, q_i \middle| R_{-i}\right)\right]$$

Since we $\textsc{pay}_i\left(r_i, q_i \middle| r_{-i}\right)$ is a strictly proper scoring rule, we have for any $i \neq t$ that $R_i = q_i$with probability 1. For t we hence have

$$\textsc{eutil}_t\left(s, p_t, y_t, c_t\right) = \mathbb{E}_{R_t \sim s_t}\left[\textsc{pay}_t\left(R_t, q_t \middle| q_{-t}\right) + c_t \cdot \mathrm{F}\left(R_t, q_{-t}\right)\right]$$

and hence

$$R_t \in \arg\max_{r_t \in \mathcal{R}_t}\left(\textsc{pay}_t\left(r_t, q_t \middle| q_{-t}\right) + c_t \cdot \mathrm{F}(r_t, q_{-t})\right).$$

In summary, for the prior D as defined above, a strategy profile is a BNE iff it guarantees that R_t as above and $R_{-t} = q_{-t}$. From Lemma 7, we know that there is a $x \in [0,1]$ such that

$$\textsc{pay}_t(q_t, q_t | q_{-t}) + c_t \cdot \mathrm{F}\left(q\right) + \epsilon \leq \textsc{pay}_t(x, q_t | q_{-t}) + c_t \cdot \mathrm{F}\left(x, q_{-t}\right)$$

with

$$\textsc{bound} := \frac{\gamma \cdot \Lambda_{\mathrm{F}}(k, \delta)^2}{8 \cdot k \cdot \delta \cdot \frac{\beta}{\gamma} \cdot H(k) + 4 \cdot k \cdot \Lambda_{\mathrm{F}}(k, \delta)}.$$

Hence, it must be the case that

$$\begin{aligned}
\textsc{pay}_t(q_t, q_t|q_{-t}) + c_t \cdot \mathrm{F}(q) + \textsc{bound} &\leq \textsc{pay}_t\left(R_t, q_t\middle|q_{-t}\right) + c_t \cdot \mathrm{F}(R_t, q_{-t}) \\
&\Updownarrow \\
\textsc{pay}_t(q_t, q_t|q_{-t}) - \textsc{pay}_t\left(R_t, q_t\middle|q_{-t}\right) + \textsc{bound} &\leq c_t \cdot (\mathrm{F}(R_t, q_{-t}) - \mathrm{F}(q)) \\
&\Downarrow \\
\textsc{bound} &\leq c_t \cdot (\mathrm{F}(R_t, q_{-t}) - \mathrm{F}(q)) \\
&\Downarrow \\
\frac{1}{\gamma} \cdot \textsc{bound} &\leq |\mathrm{F}(R_t, q_{-t}) - \mathrm{F}(q)| \\
&\Downarrow (R_{-t} = q_{-t}) \text{ and } Q = q) \\
\underbrace{\frac{\Lambda_\mathrm{F}(k,\delta)^2}{8 \cdot k \cdot \delta \cdot \frac{\beta}{\gamma} \cdot H(k) + 4 \cdot k \cdot \Lambda_\mathrm{F}(k,\delta)}}_{=:2\cdot\epsilon} &\leq |\mathrm{F}(R) - \mathrm{F}(Q)|\,.
\end{aligned}$$

Suppose that F is such that

$$\max_q |f(q) - \mathrm{F}(q)| \geq \epsilon.$$

Since any $Q = q$ is possible, and in the absence of competing incentives, we have $R = q$, hence any such F will satisfy the bound $|f(R) - f(Q)| \geq \epsilon$. Now assume

$$\max_q |f(q) - \mathrm{F}(q)| < \epsilon.$$

Then, we have for any q, r

$$\begin{aligned}
|f(q) - \mathrm{F}(r)| &\geq |\mathrm{F}(q) - \mathrm{F}(r) + f(q) - \mathrm{F}(q)| \\
&\geq |\mathrm{F}(q) - \mathrm{F}(r)| - |f(q) - \mathrm{F}(q)| \\
&\geq |\mathrm{F}(q) - \mathrm{F}(r)| - \epsilon.
\end{aligned}$$

hence, we have

$$\begin{aligned}
\Lambda_\mathrm{F}(k,\delta) :&= \max_{\mathcal{W} \in \textsc{walks}(k,\delta)} \mathbb{E}_{X\sim\mathcal{W}}\left[\sum_{i=1}^{k} \left|\mathrm{F}(X^i) - \mathrm{F}(X^{i-1})\right|\right] \\
&\geq \max_{\mathcal{W} \in \textsc{walks}(k,\delta)} \mathbb{E}_{X\sim\mathcal{W}}\left[\sum_{i=1}^{k} \left|f(X^i) - f(X^{i-1})\right| - 2\epsilon k\right] \\
&= \Lambda_f(k,\delta) - 2\epsilon k.
\end{aligned}$$

and hence

$$
\begin{aligned}
2\epsilon &\geq \frac{\Lambda_{\mathrm{F}}(k,\delta)^2}{8\cdot k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+4\cdot k\cdot\Lambda_{\mathrm{F}}(k,\delta)} \\
&\geq \frac{(\Lambda_f(k,\delta)-2\epsilon k)^2}{8\cdot k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+4\cdot k\cdot(\Lambda_f(k,\delta)-2\epsilon k)} \\
&\geq \underbrace{\frac{\Lambda_f(k,\delta)^2-2\cdot\Lambda_f(k,\delta)\cdot\epsilon k}{8\cdot k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+4\cdot k\cdot\Lambda_f(k,\delta)}}_{=:x} \\
&\Downarrow \\
2\cdot\epsilon\cdot x+2\cdot\Lambda_f(k,\delta)\cdot\epsilon k &\geq \Lambda_f(k,\delta)^2 \\
&\Downarrow \\
\epsilon &\geq \frac{\Lambda_f(k,\delta)^2}{2\cdot x+2\cdot\Lambda_f(k,\delta)\cdot k} \\
&= \frac{\Lambda_f(k,\delta)^2}{16\cdot k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+8\cdot k\cdot\Lambda_f(k,\delta)+2\cdot\Lambda_f(k,\delta)\cdot k} \\
&= \frac{\Lambda_f(k,\delta)^2}{16\cdot k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+10\cdot k\cdot\Lambda_f(k,\delta)} \\
&\geq \frac{1}{16}\frac{\Lambda_f(k,\delta)^2}{k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+k\cdot\Lambda_f(k,\delta)}
\end{aligned}
$$

Hence

$$
\begin{aligned}
|f(Q)-\mathrm{F}(R)| &\geq |\mathrm{F}(Q)-\mathrm{F}(R)|-\epsilon \\
&\geq \epsilon \\
&\geq \frac{1}{16}\frac{\Lambda_f(k,\delta)^2}{k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot H(k)+k\cdot\Lambda_f(k,\delta)} \\
&\geq \frac{1}{16}\frac{\Lambda_f(k,\delta)^2}{k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot\log(3\cdot k)+k\cdot\Lambda_f(k,\delta)}
\end{aligned}
$$

which concludes the proof of the first statement in the theorem. The second statement follows from

$$\begin{aligned}
&\frac{1}{16}\cdot\frac{\gamma}{\beta}\cdot\Delta_f - \frac{1}{16}\frac{\Lambda_f(k,\delta)^2}{k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot\log(3\cdot k)+k\cdot\Lambda_f(k,\delta)}\\
&= \frac{1}{16}\cdot\frac{\gamma}{\beta}\cdot\frac{\Lambda_f(k,\delta)^2}{\delta\cdot k\cdot\log(3\cdot k)} - \frac{1}{16}\frac{\Lambda_f(k,\delta)^2}{k\cdot\delta\cdot\frac{\beta}{\gamma}\cdot\log(3\cdot k)+k\cdot\Lambda_f(k,\delta)}\\
&= \frac{\Lambda_f(k,\delta)^2}{16}\cdot\frac{\gamma}{\beta}\cdot\left(\frac{1}{\delta\cdot k\cdot\log(3\cdot k)} - \frac{1}{k\cdot\delta\cdot\log(3\cdot k)+\frac{\gamma}{\beta}\cdot k\cdot\Lambda_f(k,\delta)}\right)\\
&= \frac{\Lambda_f(k,\delta)^2}{16}\cdot\frac{\gamma^2}{\beta^2}\cdot\left(\frac{\frac{\Lambda(k,\delta)}{\delta\cdot\log(3\cdot k)}}{\delta\cdot\log(3\cdot k)+\frac{\gamma}{\beta}\cdot\Lambda_f(k,\delta)}\right)\\
&= O(\gamma^2) \quad \text{as } \gamma\to 0.
\end{aligned}$$

□

C.2 Proof of Lemma 2

Proof. For a given $k \le m, \delta \le \sigma$, consider the random walk with
$\mathcal{K} \subseteq \mathcal{M}$ some set of size k
$X^0 = \mu$ and

$$Z_i = \begin{cases} \sim \mathcal{U}(-\delta,\delta) & \text{if } i\in\mathcal{K} \\ 0 & \text{else} \end{cases}.$$

We have

$$\begin{aligned}
\Lambda_f(k,\delta) &\ge \mathbb{E}\left[\textsc{weight}(X)\right]\\
&= \sum_{j=1}^{k}\mathbb{E}\left[\left|f(X^j)-f(X^{j-1})\right|\right]\\
&\ge \sum_{j=1}^{k}\mathbb{E}\left[\left|f(X^j)-f(X^{j-1})\right|\cdot\left[\!\left[X^j\in\mathcal{L}\text{ and }X^{j-1}\in\mathcal{L}\right]\!\right]\right]\\
&\ge \delta\cdot L\cdot\sum_{j=1}^{k}\mathbb{E}\left[\left[\!\left[X^j\in\mathcal{L},X^{j-1}\in\mathcal{L}\right]\!\right]\right].
\end{aligned}$$

Note that we have

$$X_i^j = \begin{cases} \mu+Z_i & \text{if } i\in\{\Theta^1,..,\Theta^j\} \\ \mu & \text{else.} \end{cases}$$

Since $Z_i\in[-\delta,\delta]\subseteq[-\sigma,\sigma]$ (since $\delta\le\sigma$), the condition $\left\|X^j-\mu\right\|_\infty \le \sigma$ is always satisfied. Hence

$$\mathbb{E}\left[\left[\!\left[X^j\in\mathcal{L},X^{j-1}\in\mathcal{L}\right]\!\right]\right] = \mathbb{E}\left[\left[\!\left[\left|\frac{1}{m}\sum_{i=1}^{m}X_i^j-\mu\right|\le\epsilon\text{ amd }\left|\frac{1}{m}\sum_{i=1}^{m}X_i^{j-1}-\mu\right|\le\epsilon\right]\!\right]\right].$$

Since

$$\frac{1}{m}\sum_{i=1}^{m}X_i^j = \mu+\frac{1}{m}\sum_{m=1}^{j}Z_{\Theta^m}$$

we have

$$\begin{aligned}
\mathbb{E}\left[\left[X^j \in \mathcal{L}, X^{j-1} \in \mathcal{L}\right]\right] &= \mathbb{E}\left[\left[\left|\sum_{m=1}^{j} Z_{\Theta^m}\right| \leq m \cdot \epsilon \text{ and } \left|\sum_{m=1}^{j-1} Z_{\Theta^m}\right| \leq m \cdot \epsilon\right]\right] \\
&= \mathbb{P}\left[\left|\sum_{m=1}^{j} Z_{\Theta^m}\right| \leq m \cdot \epsilon \text{ and } \left|\sum_{m=1}^{j-1} Z_{\Theta^m}\right| \leq m \cdot \epsilon\right] \\
&= 1 - \mathbb{P}\left[\left|\sum_{m=1}^{j} Z_{\Theta^m}\right| > m \cdot \epsilon \text{ or } \left|\sum_{m=1}^{j-1} Z_{\Theta^m}\right| > m \cdot \epsilon\right] \\
&\geq 1 - \left(\mathbb{P}\left[\left|\sum_{m=1}^{j} Z_{\Theta^m}\right| > m \cdot \epsilon\right] + \mathbb{P}\left[\left|\sum_{m=1}^{j} Z_{\Theta^m}\right| > m \cdot \epsilon\right]\right)
\end{aligned}$$

where the second inequality follows from the union bound. Note that $Z_{\Theta^1}, .., Z_{\Theta^1}$ are independent random variables in $-\delta, \delta$, we can hence bound these events using Hoeffding's inequality:

$$\mathbb{P}\left[\left|\sum_{m=1}^{j} Z_{\Theta^m}\right| > m \cdot \epsilon\right] \leq 2 \cdot \exp\left(-2 \cdot \frac{m^2 \cdot \epsilon^2}{j \cdot \delta^2}\right)$$

hence

$$\mathbb{E}\left[\left[X^j \in \mathcal{L}, X^{j-1} \in \mathcal{L}\right]\right] \geq 1 - 4 \cdot \exp\left(-2 \cdot \frac{m^2 \cdot \epsilon^2}{j \cdot \delta^2}\right)$$

and

$$\begin{aligned}
\Lambda_f(k, \delta) &\geq \delta \cdot L \cdot \sum_{j=1}^{k} \mathbb{E}\left[\left[X^j \in \mathcal{L}, X^{j-1} \in \mathcal{L}\right]\right] \\
&\geq \delta \cdot L \cdot \sum_{j=1}^{k} \left(1 - 4 \cdot \exp\left(-2 \cdot \frac{m^2 \cdot \epsilon^2}{j \cdot \delta^2}\right)\right) \\
&\geq \delta \cdot L \cdot k \cdot \left(1 - 4 \cdot \exp\left(-2 \cdot \frac{m^2 \cdot \epsilon^2}{k \cdot \delta^2}\right)\right).
\end{aligned}$$

Hence, recalling Definition 7, we have

$$\begin{aligned}
\Delta_f &:= \max_{k \in [n], \delta \in [0,1]} \frac{\Lambda_f(k, \delta)^2}{\delta \cdot k \cdot \log(3 \cdot k)} \\
&\geq \max_{k \leq m, \delta \leq \sigma} \frac{\Lambda_f(k, \delta)^2}{\delta \cdot k \cdot \log(3 \cdot k)}. \\
&\geq \frac{1}{\log(3 \cdot m)} \cdot \max_{k \leq m, \delta \leq \sigma} \left(k^{-1} \cdot \delta^{-1} \cdot \delta^2 \cdot L^2 \cdot k^2 \cdot \left(1 - 4 \cdot \exp\left(-2 \cdot \frac{m^2 \cdot \epsilon^2}{k \cdot \delta^2}\right)\right)^2\right) \\
&= \frac{1}{\log(3 \cdot m)} \cdot \max_{k \leq m, \delta \leq \sigma} \left(\delta \cdot L^2 \cdot k \cdot \left(1 - 4 \cdot \exp\left(-2 \cdot \frac{m^2 \cdot \epsilon^2}{k \cdot \delta^2}\right)\right)^2\right) \\
&\geq \frac{1}{\log(3 \cdot m)} \cdot \max_{\delta \leq \sigma} \left(\delta \cdot L^2 \cdot m \cdot \left(1 - 4 \cdot \exp\left(-2 \cdot \frac{m \cdot \epsilon^2}{\delta^2}\right)\right)^2\right)
\end{aligned}$$

With

$$\delta = \min\left\{\frac{m \cdot \epsilon}{\sqrt{m} \cdot 2}, \sigma\right\}$$

we have

$$\begin{aligned}
\Delta_f &\geq \frac{1}{\log(3 \cdot m)} \cdot \min\left\{\frac{m \cdot \epsilon}{\sqrt{m} \cdot 2}, \sigma\right\} \cdot L^2 \cdot m \cdot (1 - 4 \cdot \exp(-8))^2 \\
&= \frac{1}{\log(3 \cdot m)} \cdot \min\left\{\frac{m \cdot \epsilon}{\sqrt{m} \cdot 2}, \sigma\right\} \cdot L^2 \cdot m \cdot \frac{2}{3} \\
&\geq \frac{2}{3} \cdot \frac{1}{\log(3 \cdot m)} \cdot \sigma \cdot m \cdot L^2 \cdot \min\left\{\frac{\sqrt{m} \cdot \epsilon}{2 \cdot \sigma}, 1\right\} \\
&\geq \frac{1}{3} \cdot \frac{1}{\log(3 \cdot m)} \cdot \sigma \cdot m \cdot L^2 \cdot \min\left\{\frac{\sqrt{m} \cdot \epsilon}{\sigma}, 1\right\}.
\end{aligned}$$

□

C.3 Proof of Theorem 7

Proof. For convenience, we will sometimes use the notation

$$\textsc{pay}_i(r_i, q_i | r_{-i}) := \mathbb{E}_{O \sim q_i}\left[\textsc{pay}_i(r, O)\right].$$

Note that $\textsc{pay}_i(r_i, o | r_{-i}) = \textsc{pay}_i(r, o)\ \forall o \in \mathcal{O}$. Recall recommender i's utility

$$\begin{aligned}
\textsc{util}_i\left(p_i, y, c_i, r\right) &= \mathbb{E}_{O \sim p_i(\cdot|y)}\left[\textsc{pay}_i(r, O)\right] + c_i \cdot \mathrm{F}(r) \\
\text{(using ADA-PAYDefinition 8)} &= \mathbb{E}_{O \sim p_i(\cdot|y)}\left[\textsc{ada-pay}_i(r_i, O)\right] + c_i \cdot f(r) \\
\text{(due to linearity)} &= \textsc{ada-pay}_i\left(r_i, p_i(O = 1|y)\right) + c_i \cdot f(r)
\end{aligned}$$

and hence

$$\begin{aligned}
\textsc{eutil}_i\left(s, p_i, y_i, c_i\right) &= \mathbb{E}\left[\textsc{util}_i\left(P_i, Y, C_i, R\right) \middle| s, p_i, y_i, c_i\right] \\
&= \mathbb{E}\left[\textsc{ada-pay}_i\left(R_i, p_i(O = 1|Y)\right) + c_i \cdot f(R) \middle| s, p_i, y_i, c_i\right] \\
&= \mathbb{E}\left[\textsc{ada-pay}_i\left(R_i, p_i(O = 1|y_i)\right) + c_i \cdot f(R) \middle| s, p_i, y_i, c_i\right] \\
&= \mathbb{E}\left[\textsc{ada-pay}_i\left(R_i, q_i\right) + c_i \cdot f(R) \middle| s, p_i, y_i, c_i\right]
\end{aligned}$$

with $q_i := p_i(O = 1|y_i)$ the true belief. Clearly, any rationalizable strategy s_i will play only r_i that are a best response. Hence, in the setting of the theorem, there must exist a generalized derivative ∂ such that we have

$$
\begin{aligned}
0 &= \partial_{r_i}\mathbb{E}\left[\text{ADA-PAY}_i\left(r_i, q_i\right) + c_i \cdot f(r_i, R_{-i})\middle| s_{-i}, p_i, y_i, c_i\right] \\
&\Updownarrow \\
-\partial_{r_i}\text{ADA-PAY}_i\left(r_i, q_i\right) &= c_i \cdot \mathbb{E}\left[\partial_{r_i} f(r_i, R_{-i})\middle| s_{-i}, p_i, y_i, c_i\right] \\
&\Updownarrow \\
-\frac{\beta}{3\cdot|\mathcal{M}|}\cdot\frac{\delta_i(r_i)(q_i - r_i)}{\int_0^1 \delta_i(x)\cdot dx} &= c_i \cdot \mathbb{E}\left[\partial_{r_i} f(r_i, R_{-i})\middle| s_{-i}, p_i, y_i, c_i\right] \\
&\Updownarrow \\
\frac{\beta}{3\cdot|\mathcal{M}|}\cdot\frac{(r_i - q_i)}{\int_0^1 \delta_i(x)\cdot dx} &= c_i \cdot \mathbb{E}\left[\frac{\partial_{r_i} f(r_i, R_{-i})}{\delta_i(r_i)}\middle| s_{-i}, p_i, y_i, c_i\right] \\
&\Updownarrow \\
r_i - q_i &= \frac{3\cdot|\mathcal{M}|\cdot c_i}{\beta}\cdot\mathbb{E}\left[\frac{\partial_{r_i} f(r_i, R_{-i})}{\delta_i(r_i)}\middle| s_{-i}, p_i, y_i, c_i\right]\cdot\int_0^1 \delta_i(x)\cdot dx \\
&\Updownarrow \\
|r_i - q_i| &= \left|\frac{3\cdot|\mathcal{M}|\cdot c_i}{\beta}\mathbb{E}\left[\frac{\partial_{r_i} f(r_i, R_{-i})}{\delta_i(r_i)}\middle| s_{-i}, p_i, y_i, c_i\right]\int_0^1 \delta_i(x)\cdot dx\right| \\
&\leq \frac{3\cdot|\mathcal{M}|\cdot|c_i|}{\beta}\left(\int_0^1 \delta_i(x)\cdot dx\right) \\
&= \frac{3\cdot|\mathcal{M}|\cdot|c_i|}{\beta}\left(2\cdot\int_0^1 \max_{r_{-i}\in\mathcal{R}_{-i}}\left|\partial_{r_i} f(r)\right| dr_i\right).
\end{aligned}
$$

This holds regardless of others' reports. Hence, we have

$$
\begin{aligned}
|f(R) - f(Q_i, R_{-i})| &\leq \frac{6\cdot|c_i|}{\beta}\left(|\mathcal{M}|\cdot\int_0^1 \max_{r_{-i}\in\mathcal{R}_{-i}}\left|\partial_{r_i} f(r)\right| dr_i\right)\cdot\left(\max_{r\in\mathcal{R}}\left|\partial_{r_i} f(r)\right|\right) \\
&\leq \frac{6\cdot|c_i|}{\beta}\cdot\nabla_f.
\end{aligned}
$$

Hence, for the total deviation, we have

$$
\begin{aligned}
|\mathrm{F}(R) - f(Q)| &= |f(R) - f(Q)| \\
&\leq \frac{6\cdot\sum_{i\in\mathcal{M}}|c_i|}{\beta}\cdot\nabla_f \\
&\leq \frac{6\cdot\gamma}{\beta}\cdot\nabla_f.
\end{aligned}
$$

□

C.4 Proof of Theorem 9

Proof. Since here, we are trying to ensure that $\mathrm{F}(R) = f(Q)$, the mechanism must choose $\mathrm{F} = f$, we will hence use them interchangeably in this proof. Now, fix some $k \in [n]$,$\delta \in [0, 1]$, and let q, t,

x, as in Lemma 7. Suppose that the prior D only has support at

$$\begin{aligned} P_i(o|y_i) = p_i(o|y_i) &:= \begin{cases} q_i & \text{if } o = 1 \\ 1 - q_i & \text{if } o = 0 \end{cases} \quad \forall y_i \\ V &= v = v \\ W &= \{t\} \\ Y &= y := 0. \end{aligned}$$

There is no uncertainty in types, hence it is not an incomplete information game. We can hence identify the briber strategy s^{BRI} with a distribution over C_t, since it is known that $C_{-t} = 0$ (due to $W = \{t\}$), and recommender $i \neq t$'s strategy s_i with a distribution over report R_i, and recommender t's strategy $s_t(c)$ with a distribution over R_t for each bribe c_t. Recommender i's utility is

$$\begin{aligned} \textsc{util}_i\left(p_i, y, c_i, r\right) &:= \mathbb{E}_{O \sim p_i(\cdot|y)}\left[\textsc{pay}_i(r, O) + c_i \cdot \mathrm{F}(r)\right] \\ &= \textsc{pay}_i(r_i, q_i | r_{-i}) + c_i \cdot \mathrm{F}(r). \end{aligned}$$

Clearly, since $\textsc{pay}_i(r_i, q_i|r_{-i})$ is a strictly proper scoring rule, any PBE must satisfy $R_{-t} = q_{-t}$. For recommender t, we have

$$\begin{aligned} \textsc{eutil}_t\left(s, s^{\text{BRI}}, p_t, y_t, c_t\right) : &= \mathbb{E}\left[\textsc{util}_t\left(P_t, Y, C_t, R\right) \middle| s, s^{\text{BRI}}, p_t, y_t, c_t\right] \\ &= \mathbb{E}\left[\textsc{util}_t\left(p_t, y, c_t, (R_t, q_{-t})\right) \middle| s, s^{\text{BRI}}, p_t, y_t, c_t\right] \\ &= \mathbb{E}_{R_t \sim s_t(c_t)}\left[\textsc{util}_t\left(p_t, y, c_t, (R_t, q_{-t})\right)\right] \\ &= \mathbb{E}_{R_t \sim s_t(c_t)}\left[\textsc{pay}_t\left(R_t, q_t \middle| q_{-t}\right) + c_t \cdot \mathrm{F}\left(R_t, q_{-t}\right)\right]. \end{aligned}$$

A PBE (Definition 10) requires that R_t be optimal for any C_t, hence

$$R_t \in \arg\max_{r_t \in \mathcal{R}_t}\left(\textsc{pay}_t\left(r_t, q_t \middle| q_{-t}\right) + C_t \cdot \mathrm{F}(r_t, q_{-t})\right).$$

From Lemma 7, we know that there is a $x \in [0, 1]$ such that

$$\textsc{pay}_t(q_t, q_t|q_{-t}) + c_t \cdot \mathrm{F}\left(q\right) + \textsc{bound} \leq \textsc{pay}_t(x, q_t|q_{-t}) + c_t \cdot \mathrm{F}\left(x, q_{-t}\right)$$

with

$$\textsc{bound} := \frac{|c_t| \cdot \Lambda_{\mathrm{F}}(k, \delta)^2}{8 \cdot k \cdot \delta \cdot \frac{\beta}{|c_t|} \cdot H(k) + 4 \cdot k \cdot \Lambda_{\mathrm{F}}(k, \delta)}.$$

Hence, it must be the case that

$$\begin{aligned}
\textsc{pay}_t(q_t, q_t|q_{-t}) + c_t \cdot \mathrm{F}(q) + \textsc{bound} &\leq \textsc{pay}_t\left(R_t, q_t\middle|q_{-t}\right) + c_t \cdot \mathrm{F}(R_t, q_{-t}) \\
&\Updownarrow \\
\textsc{pay}_t(q_t, q_t|q_{-t}) - \textsc{pay}_t\left(R_t, q_t\middle|q_{-t}\right) + \textsc{bound} &\leq c_t \cdot (\mathrm{F}(R_t, q_{-t}) - \mathrm{F}(q)) \\
&\Downarrow \\
\textsc{bound} &\leq c_t \cdot (\mathrm{F}(R_t, q_{-t}) - \mathrm{F}(q)) \\
&\Downarrow \\
\frac{1}{\gamma} \cdot \textsc{bound} &\leq |\mathrm{F}(R_t, q_{-t}) - \mathrm{F}(q)| \\
&\Downarrow (R_{-t} = q_{-t} \text{ and } Q = q) \\
\frac{\Lambda_{\mathrm{F}}(k,\delta)^2}{8 \cdot k \cdot \delta \cdot \frac{\beta}{|C_t|} \cdot H(k) + 4 \cdot k \cdot \Lambda_{\mathrm{F}}(k,\delta)} &\leq |\mathrm{F}(R) - \mathrm{F}(Q)|\,.
\end{aligned}$$

For the briber, we have

$$\begin{aligned}
\textsc{eutil}^{\textsc{bri}}(s, s^{\textsc{bri}}, v, w) &:= \mathbb{E}\left[\textsc{util}^{\textsc{bri}}(V, W, C, R)\,\middle|s, s^{\textsc{bri}}, v, w\right] \\
&= \mathbb{E}_{\hat{S}_t \sim s_t, C_t \sim s^{\textsc{bri}}}\left[(v - |C_t|) \cdot \mathrm{F}(\hat{S}_t(C_t), q_{-t})\right].
\end{aligned}$$

As we have seen above, the recommenders' PBE strategy profile does not depend on the briber's type or strategy, for the prior D defined above, but only on the bribe they observe. Hence, a strategy profile $s, s^{\textsc{bri}}$ is a PBE iff the recommender strategy profile s is such that $R_{-t} = q_{-t}$ and R_t is as above, along with a briber strategy profile $s^{\textsc{bri}}$ being such that

$$C_t \in \arg\max_{c_t \in \mathbb{R}} \mathbb{E}_{\hat{S}_t \sim s_t}\left[(v - |c_t|) \cdot \mathrm{F}(\hat{S}_t(c_t), q_{-t})\right].$$

A necessary condition for the briber to not bribe is hence that for all $c_t \in \mathbb{R}, |c_t| > 0$ we have

$$v \cdot \mathrm{F}(q) \geq \mathbb{E}_{\hat{S}_t \sim s_t}\left[(v - |c_t|) \cdot \mathrm{F}(\hat{S}_t(c_t), q_{-t})\right].$$

We know from above that

$$\begin{aligned}
v \cdot \mathrm{F}(q) &\geq \mathbb{E}_{\hat{S}_t \sim s_t}\left[(v - |c_t|) \cdot \mathrm{F}(\hat{S}_t(c_t), q_{-t})\right] \\
&\Updownarrow \\
0 &\geq (v - |c_t|) \cdot \mathbb{E}_{\hat{S}_t \sim s_t}\left[\mathrm{F}(\hat{S}_t(c_t), q_{-t}) - \mathrm{F}(q)\right] - |c_t| \cdot \mathrm{F}(q) \\
&\Downarrow \\
0 &\geq (v - |c_t|) \cdot \mathbb{E}_{\hat{S}_t \sim s_t}\left[\mathrm{F}(\hat{S}_t(c_t), q_{-t}) - \mathrm{F}(q)\right] - |c_t| \cdot \max_{q \in Q} \mathrm{F}(q) \\
&\Downarrow \\
0 &\geq (v - |c_t|) \frac{\Lambda_{\mathrm{F}}(k,\delta)^2}{8 \cdot k \cdot \delta \cdot \frac{\beta}{|c_t|} \cdot H(k) + 4 \cdot k \cdot \Lambda_{\mathrm{F}}(k,\delta)} - |c_t| \cdot \max_{q \in Q} \mathrm{F}(q) \\
&\Downarrow \\
0 &\geq (v - |c_t|) \frac{\Lambda_{\mathrm{F}}(k,\delta)^2}{8 \cdot k \cdot \delta \cdot \beta \cdot H(k) + 4 \cdot k \cdot \Lambda_{\mathrm{F}}(k,\delta) \cdot |c_t|} - \max_{q \in Q} \mathrm{F}(q) \\
&\Downarrow \\
0 &\geq (v - |c_t|) \frac{\Lambda_{\mathrm{F}}(k,\delta)^2}{8 \cdot k \cdot \delta \cdot \beta \cdot \log(3k) + 4 \cdot k \cdot \Lambda_{\mathrm{F}}(k,\delta) \cdot |c_t|} - \max_{q \in Q} \mathrm{F}(q)
\end{aligned}$$

which is violated if

$$\frac{1}{8 \cdot \max_{q \in Q} \mathrm{F}(q)} \cdot \frac{\Lambda_{\mathrm{F}}(k,\delta)^2}{k \cdot \delta \cdot \log(3k)} > \frac{\beta}{v}.$$

□

C.5 Proof of Theorem 10

Proof. Along similar lines as in the proof in Appendix C.3, we obtain, using the adaptive payment mechanism (Definition 8), that

$$|\mathrm{F}(R) - \mathrm{F}(Q)| \leq \frac{6 \cdot \sum_{i \in [n]} |C_i|}{\beta} \cdot \nabla_f$$

with probability 1. Using $\bar{C} := \sum_{i \in [n]} |C_i|$, the utility the briber gains is at most

$$\begin{aligned}
\left(V - \bar{C}\right) \cdot \mathrm{F}(R) - V \cdot \mathrm{F}(Q) &= \left(V - \bar{C}\right) \cdot (\mathrm{F}(R) - \mathrm{F}(Q)) - \bar{C} \cdot \mathrm{F}(Q) \\
&\leq \left(V - \bar{C}\right) \cdot \frac{6 \cdot \bar{C}}{\beta} \cdot \nabla_f - \bar{C} \cdot \min_{q \in Q} f(q).
\end{aligned}$$

Hence, if for all $\bar{C} > 0$ we have

$$0 > \left(V - \bar{C}\right) \cdot \frac{6 \cdot \bar{C}}{\beta} \cdot \nabla_f - \bar{C} \cdot \min_{q \in Q} f(q),$$

then there will be no bribe. A sufficient condition is

$$
\begin{aligned}
&\Uparrow \\
0 &> \left(V - \bar{C}\right) \cdot \frac{6}{\beta} \cdot \nabla_f - \min_{q \in Q} f(q) \\
&\Uparrow \\
0 &\geq 6 \cdot \frac{v}{\beta} \cdot \nabla_f - \min_{q \in Q} f(q).
\end{aligned}
$$

□