Buying Private Data without Verification

Arpita Ghosh[*]   Katrina Ligett[†]   Aaron Roth[‡]   Grant Schoenebeck[§]

April 25, 2014

Abstract

We consider the problem of designing a survey to aggregate non-verifiable information from a privacy-sensitive population: an analyst wants to compute some aggregate statistic from the private bits held by each member of a population, but cannot verify the correctness of the bits reported by participants in his survey. Individuals in the population are strategic agents with a cost for privacy, *i.e.*, they not only account for the payments they expect to receive from the mechanism, but also their privacy costs from any information revealed about them by the mechanism's outcome—the computed statistic as well as the payments—to determine their utilities. How can the analyst design payments to obtain an accurate estimate of the population statistic when individuals strategically decide *both* whether to participate and whether to truthfully report their sensitive information?

We design a differentially private peer-prediction mechanism [MRZ05] that supports *accurate* estimation of the population statistic as a Bayes-Nash equilibrium in settings where agents have explicit preferences for privacy. The mechanism requires knowledge of the marginal prior distribution on bits b_i, but does not need full knowledge of the marginal distribution on the costs c_i, instead requiring only an approximate upper bound. Our mechanism guarantees ϵ-differential privacy to each agent i against any adversary who can observe the statistical estimate output by the mechanism, as well as the payments made to the $n - 1$ other agents $j \neq i$. Finally, we show that with slightly more structured assumptions on the privacy cost functions of each agent [CCK+13], the cost of running the survey goes to 0 as the number of agents diverges.

1 Introduction

Consider the problem faced by a researcher who would like to compute some (unknown) statistic about a target population: for example, the prevalence of AIDS among university professors. He could, of course, run a survey and 'just ask'; the obstacle he faces is that an individual's AIDS status is sensitive data, and the individuals being surveyed may be concerned that some harm might befall them if they were to participate in such a survey. This is a problem for the researcher, since individuals are not obligated to participate in his survey—to solve this problem, he might offer a (possibly different) payment to each participant to compensate them for such concerns and offset their 'privacy' costs from participating in his survey. The researcher still has a problem, though—even if he manages to recruit individuals to participate in his survey

---

[*]Cornell University

[†]Caltech. Supported in part by an NSF CAREER award (CNS-1254169), the US-Israel Binational Science Foundation (grant 2012348), the Charles Lee Powell Foundation, a Google Faculty Research Award, and a Microsoft Faculty Fellowship.

[‡]University of Pennsylvania. Supported in part by an NSF CAREER award, under NSF grants CCF-1101389 and CNS-1065060, and a Google Focused Research Award.

[§]University of Michigan

(for instance, incentivized by the payment offered for participation), these participants are still not obligated to truthfully report their private data, *nor* does our researcher have any direct means to verify the truth of their responses.

The "sensitive surveyor's problem" [GR13] attempts to model the problem of procuring data from individuals, each with a cost of privacy, in order to estimate some aggregate statistic of a population, and has since been studied in some depth [GR13, FL12, RS12, LR12, GL13, NVX14]. This stream of work, however, has thus far only modeled settings in which the data collected from each participant are *verifiable*, and the only information that a participant can misreport is her privacy cost determining the compensation she requires for participation—that is, an individual may choose to not participate in the survey, but if participating, she cannot lie about her data.[1] As evident from our cartoon AIDS survey example, however, there are clearly also scenarios where a researcher *cannot* directly verify the truth of the responses to his survey. A researcher in such a scenario faces an additional problem in computing a reliable estimate of an aggregate statistic beyond cheaply eliciting participation—recruiting a large enough population of respondents does not automatically lead to representative estimates because individuals may still lie about their data, compromising the accuracy of the estimate. How can payments be designed to ensure that the survey produces an accurate estimate of the population statistic, when individuals strategically decide both whether or not to participate *and* whether or not to truthfully report their sensitive data, accounting both for the payments they expect to receive from the mechanism and their privacy costs from the mechanism's outcome?

The problem of eliciting unverifiable[2] information from strategic agents is addressed by the peer prediction mechanism [MRZ05], which uses proper scoring rules to design payments such that truthful reporting is a (Bayes)-Nash equilibrium of the mechanism. Agents in [MRZ05], however, have no concern for privacy—they do not care about what the outcomes computed by the mechanism might reveal about their reports, and act only to maximize their expected payment from the mechanism. The peer-prediction mechanism, therefore, does not directly address the following problem of eliciting information from privacy-sensitive individuals.

Consider an abstract model of such a privacy-sensitive population: each agent in the population has a private (sensitive) bit b_i, as well as a parameter c_i which bounds her cost for privacy as follows. If the mechanism that agents interact with is ϵ-differentially private [DMNS06], then an agent's cost (which is potentially distinct for each outcome) is at most, though not necessarily exactly, ϵc_i. The bit-cost pairs (b_i, c_i) for each individual are drawn from a (not necessarily independent) commonly known prior distribution. Implementing a peer prediction mechanism in such a setting faces two main obstacles:

1. The way that the peer prediction mechanism incentivizes truth-telling despite unverifiability is by rewarding a participant's stated bit's correlation with that of another participant. A participating agent i in the peer prediction mechanism of [MRZ05] is paired with a uniformly randomly selected agent j, and paid as a deterministic function of her own reported bit $\hat{b}_i$, as well as the reported bit $\hat{b}_j$ of agent j. Such a payment rule is inherently disclosive;[3] for the payments to satisfy differential privacy, we must instead base them on *perturbed aggregates* of all players' reports. Relatedly, if the privacy costs c_i are drawn from a distribution with unbounded support, no finite level of payment will encourage full participation with truthful reporting, in contrast to [MRZ05]. Hence, the derivation of Bayes-Nash equilibrium conditions is more delicate: we require players to be strictly incentivized to report their

---

[1]For instance, her data will be automatically collected from some database once she consents to its use, or equivalently, false reports can easily be identified, in which case not compensating individuals for false reports effectively eliminates the incentive to participate and yet lie about the sensitive data.

[2]*i.e.*, where agents' reports cannot be compared against some ground truth for verification

[3]since it is computed deterministically as a function of the reported bit of a single individual j

true bit, which can be interpreted as a prediction of the average value of all other reported bits, but this incentive must be robust to the noise that is explicitly added to the computed aggregate (which determines the payment) in order to preserve privacy, as well as to the error in the estimate from lack of full participation.

2. As mentioned above, in order to guarantee differential privacy, we require that the output (and payments) we compute be perturbed, which introduces error in our estimate of the statistic that we wish to compute. To obtain some fixed level of accuracy, therefore, we must deal with a tradeoff: on the one hand, increasing the perturbation rate directly decreases accuracy by virtue of the noise we add. On the other hand, it increases privacy, which can in turn increase participation, which *increases* accuracy. In order to find an equilibrium that matches our accuracy goals, we must manage this tradeoff.

In this paper, we give a differentially private peer prediction mechanism that supports the *accurate* estimation of a population statistic as a Bayes Nash equilibrium, even in settings where agents have explicit preferences for privacy. Our mechanism, which guarantees ϵ-differential privacy to each agent i against any adversary who can observe the statistical estimate output by the mechanism as well as the payments made to the $n-1$ other agents $j \neq i$ (but not the payment made to agent i,[4]) requires knowledge of the marginal prior distribution on bits b_i, but need only know a crude upper bound on the marginal distribution on costs c_i. We show that with a slightly more specific assumption about agent privacy costs [CCK$^+$13], the mechanism's total cost of compensating players for their privacy goes to zero as the number of agents diverges, while supporting accurate estimation of the statistic in equilibrium.

1.1 Related Work

There are two main strands of related work: that in the "sensitive surveys" literature, and that in the peer prediction literature.

1.1.1 Privacy Aware Surveys

The problem of estimating a population statistic among strategic agents who have explicit costs for privacy was introduced by Ghosh and Roth [GR13], who considered the problem of designing a prior-free direct revelation mechanism for this task. They considered two settings: one in which agents' costs c_i can be correlated with their private bits b_i (and hence the costs themselves can potentially be disclosive, and computations on them can harm privacy), and another in which agents' costs are independent from their private bits (or at least agents do not regard themselves as experiencing privacy costs due to computations done only on their reported c_i values). In the first (more interesting) model, [GR13] prove an impossibility result, showing that no individually rational mechanism that makes finite payments can achieve a non-trivial estimate of the underlying population statistic. In the second (less realistic) model, [GR13] give dominant strategy truthful mechanisms for either a) producing a statistic to some target accuracy α, with optimal payments relative to an envy-free benchmark, or b) producing a maximally accurate statistic given a fixed payment budget.

---

[4]We assume that the surveyor is trusted, and that agents are concerned about maintaining privacy from the other people in the survey, as well as from external observers of the survey outcome. We think of payments made to a particular agent as confidential, so that an outside observer, together with a coalition of survey participants, could in the worst case learn the outcome of the survey, as well as the payments made to all players $j \neq i$, and use these to form inferences about agent i's private data. (Note that if the adversary could see the payment made to agent i, then it is not hard to see that almost nothing can be done to preserve privacy: observing the payment of player i reveals that her cost for participating is lower than the payment, which is a violation of differential privacy (of the costs)).

Much of the subsequent work has focused on circumventing and extending the impossibility result that [GR13] prove in the setting in which agents' bits and costs are correlated. Fleischer and Lyu [FL12] consider a Bayesian setting in which there are two known prior distributions on costs, D_0 and D_1. Agents with $b_i = 1$ have their cost drawn from D_1, and agents with $b_i = 0$ have their costs drawn from D_0. The mechanism knows the priors, but does not know the proportion of individuals with $b_i = 1$. [FL12] give a clever contracting scheme that makes the distribution on agent participation decisions independent of their bit, and from this argue that they can incentivize agent participation even in the presence of correlations between bits and costs. This mechanism, however, does require that these priors be known exactly to the mechanism: both the privacy guarantee and the truthfulness guarantee depend on the accuracy of these priors.

Ligett and Roth [LR12] in contrast give a model in which the impossibility result of [GR13] is circumvented by relaxing the individual rationality requirement. In this model, the mechanism has the power to *observe non-participation decisions*—which, when costs and bits are correlated, can result in privacy loss for the non-participating agents. However, [LR12] only require that the mechanism make (individually rational) payments to participating agents, and not to non-participating ones.

Most recently, an elegant paper of Nissim, Vadhan, and Xiao [NVX14] revisits this problem and offers two results. First, they significantly strengthen the impossibility result of [GR13] to hold under a much weaker set of assumptions. Second, they circumvent the impossibility result by making the natural assumption that agent costs are positively correlated with their bit: that is, they make the assumption that there is a clear "more sensitive" bit value, such that an agent whose bit flips from $b_i = 0$ to $b_i = 1$ will only have a higher, and never lower cost for privacy. By promising privacy only for agents that have this property, they are able to circumvent the impossibility theorem.

Common to all these papers (which are surveyed in [PR13], along with a broader set of work on the intersection of privacy and mechanism design) is the assumption that agent bits b_i are *verifiable*, and that agents only have the ability to misreport their costs c_i. This represents the main departure of the present paper from previous work on the sensitive surveyor's problem: we allow agents to mis-report their bit b_i.

1.1.2 Privacy Cost Functions

Any work that seeks to model strategic agents in the presence of privacy concerns must grapple with the problem of how to model agents' costs for privacy. *Differential privacy*, introduced by Dwork, McSherry, Nissim, and Smith [DMNS06] provides a formal measure which gives a linear upper bound on the degree to which an agent can decrease his expected future utility via his decision to participate in a mechanism (see [GR13] for further discussion of this). [GR13] use a model that assumes agent costs are *exactly* linear: that they experience cost $c_i \epsilon$ for some c_i when their data are used in an ϵ-differentially private way. Nissim, Orlandi, and Smorodinsky [NOS12] propose assuming that agents' privacy costs can be arbitrary, but *upper bounded* by some linear cost $c_i \epsilon$. This is the most general assumption in the literature, and the one that we adopt for most of our paper.

Chen et al. [CCK$^+$13] propose and justify an outcome-based measure (also based on differential privacy) which usually results in much lower (sublinear) costs, which have several nice analytic properties. We also adopt this model in Section 3.4 as an alternative (and relaxed) definition of privacy.

1.1.3 Peer Prediction

The peer-prediction method, introduced by Miller, Resnick and Zeckhauser [MRZ05], is a mechanism for truthfully eliciting information from agents in the absence of a verifiable ground truth against which to compare agents' reports. The mechanism in [MRZ05] uses proper scoring rules to reward agents for

making reports that are predictive of other agents' reports, to ensure that truthtelling is a Nash equilibrium. There has since been a stream of work on several variants of the original peer-prediction model, motivated by opinion elicitation in online settings such as reviews and reputations where there is no objective ground truth [JF06, JF07, JF09, WP11, WP12b]. The Bayesian Truth Serum (BTS) mechanism and its variants [Pre04, WP12a] use a different technique for eliciting truthful reports of unverifiable information that do not require the assumption of a common prior. Finally, there is also work focused explicitly on incentive-compatible mechanisms for surveying a population to acquire an estimate of some statistic [LS08, JF08, GRP09, PRGJ11]. The papers amongst these that are closest to our problem are perhaps [JF06], which uses automated mechanism design to minimize the budget required by an incentive-compatible information elicitation mechanism, albeit in a setting with equal participation costs, and [GRP09] which presents a truthful, weighted, budget-balanced mechanism for collective revelation.

This information elicitation literature, however, studies models that differ fundamentally from our setting in that agents derive utility only from the payment they receive from the mechanism, and not from any use of the reported information itself. Specifically, this literature thus far has focused, to the best of our knowledge, entirely on settings where agents do not derive any (dis)utility from any outputs computed from the reports made to the mechanism, unlike in our setting with privacy costs (where the outputs are the payments made to (other) agents, and the aggregate statistic computed by the analyst). Also, agents' participation costs in this literature (which in fact has largely assumed equal participation costs for all agents) do not relate to the (private) information the surveyor wishes to elicit from them. These two differences between our setting with a privacy sensitive population and the models in the information elicitation literature significantly alter both the incentives that need to be provided for truthful reporting, as well as the tradeoffs between the total cost incurred by the mechanism and the accuracy of its output.

2 Model

We now present a model for the problem of designing a survey to accurately aggregate *non-verifiable* information from a population of privacy-sensitive agents, each of whom strategically chooses whether or not to participate, and whether or not to truthfully report her private bit to the survey.

Agents. There are n agents in the population, each with a private bit $b_i \in \{0, 1\}$ of interest to an analyst, and a private cost coefficient c_i that characterizes the agent's disutility from any privacy loss from the survey. The vector of bit-cost pairs $(b, c) = [(b_1, c_1), \ldots, (b_n, c_n)]$ describing the population of n agents is drawn from a known joint distribution $\mathcal{P}$ over $\{0, 1\}^n \times \mathbb{R}_{\geq 0}^n$; since the prior $\mathcal{P}$ is a joint distribution over bit-cost *pairs*, this model allows for agents to have privacy costs c_i that are correlated with their sensitive private bit b_i. (We note that the mechanisms studied in this paper elicit agents' *bits*, but do *not* ever require reports of the *costs*—the privacy costs are only used by agents in their own utility calculations to determine their optimal action trading off the costs and benefits of participation and truthful reporting.)

We make two assumptions about the joint distribution $\mathcal{P}$. (i) We assume that $\mathcal{P}$ is *symmetric* over agents, in that all agents are 'equal' in terms of what bit-cost pair they may draw. Formally, for every permutation σ and every $b \in \{0, 1\}^n, c \in \mathbb{R}_{\geq 0}^n$, we have $\Pr_{\mathcal{P}}[(b, c)] = \Pr_{\mathcal{P}}[\sigma((b, c))]$, *i.e.*, all permutations of a given bit-cost vector describing the population are equally likely.[5] (ii) Second, we assume that an agent's *cost* c_i does not give her any information about other agents, beyond what is already conveyed to her by her own *bit* b_i. Formally, the posterior distribution on other agents' bits b_{-i} and costs c_{-i} that any agent i can compute

---

[5]Of course, note that this does not mean that all individual bit-cost *pairs* are equally likely; it only says that all *agents* are equally likely to have a particular pair.

after observing his own type (b_i, c_i) is conditionally independent of c_i given b_i: for every $b_i, b_{-i}, c_i, c'_i, c_{-i}$, $\Pr[b_{-i}|b_i, c_i] = \Pr[b_{-i}|b_i, c'_i]$. This is a natural condition that can be satisfied in a number of ways—for example, it is satisfied if agents *first* draw a bit b_i from an underlying distribution on bits, and then draw their cost from a distribution parameterized by b_i, as in the model considered by [FL12].

Since an agent's cost, conditional on her bit, does not affect the distribution of remaining agents' bits, and since agents are symmetric, the distribution over the $n-1$ remaining draws of private bits b_{-i} given the draw of b_i does not depend either on the identity i of the agent who made the draw, nor on her cost c_i. We will use the following notation for these common posterior distributions.

Definition 2.1. *[$\mathcal{P}_0, \mathcal{P}_1, \mathcal{B}, \mathcal{B}_0, \mathcal{B}_1, \mathcal{C}, \mathcal{C}_0, \mathcal{C}_1$.] Consider a prior $\mathcal{P}$ over the n-vector of bit-cost pairs $[(b_i, c_i)]$. We denote the conditional joint distribution of the vector $(b, c)_{-i} \in (\{0, 1\} \times \mathbb{R})^{n-1}$ conditioned on $b_i = 0$ by $\mathcal{P}_0$ and that conditioned on $b_i = 1$ by $\mathcal{P}_1$. We denote the marginal distribution of $\mathcal{P}$ (and $\mathcal{P}_0$ and $\mathcal{P}_1$) on the bits as $\mathcal{B}$ (and $\mathcal{B}_0$ and $\mathcal{B}_1$) and the marginal distribution of the costs as $\mathcal{C}$ (and $\mathcal{C}_0$ and $\mathcal{C}_1$)*

Mechanisms. There is an analyst who would like to acquire each agent's private bit b_i in order to perform some computation on it. The analyst uses a *mechanism* M to transform the set of bits $\hat{b}_i$ reported by participating agents into an outcome o; in exchange, the mechanism can make *payments* π_i to the agents for their reports. Formally, a mechanism is a randomized mapping $M : \{0, 1\}^n \to \mathcal{O} \times \mathbb{R}^n_{\geq 0}$, taking as input the vector of reported bits $\hat{b}_i \in \{0, 1, \bot\}$ (where $\bot$ represents the decision to decline to provide one's bits), and produces an output $M(\hat{b}) = (o, \pi)$ where $o \in \mathcal{O}$ is a publicly observed outcome in some abstract outcome space $\mathcal{O}$, and $\pi \in \mathbb{R}^n_{\geq 0}$ is a vector of payments. We will think of the analyst as interested in computing the fraction of the population that has $b_i = 1$, so that the mechanisms we design will have outcome space $\mathcal{O} = [0, 1] \subseteq \mathbb{R}$.

Unlike in prior literature, we do *not* assume that the bits b_i are verifiable by the analyst. In our setting, in addition to declining to provide their bit to the mechanism, agents can *lie* about their private bits b_i, and will do so if it improves their utilities, which we describe next.

Agent utilities. An agent's utility is the difference between the payment π_i she receives from the analyst, and the cost she incurs from privacy losses from the outcome of the mechanism, modeled as follows.

Each agent in the population is concerned about *privacy*, in the sense of what might be revealed about her sensitive bit b_i by information that becomes available to others as a result of the survey. This includes both the publicly released outcome o, as well as the payments π_{-i} made to the *other* agents by the mechanism; we assume, however, that the payment π_i made to player i is unobservable to anyone other than player i. We note that this is the strongest privacy model possible, as it is easy to see that if an adversary can see player i's payment, then no mechanism can be both finitely differentially private and truthful in our setting. We measure privacy loss in our setting using a variant of *differential privacy* [DMNS06] called *joint-differential privacy*, introduced by [KPRU14]:

Definition 2.2 (Joint Differential Privacy). *Let $M(b)_{-i} = (o, \pi_{-i})$ denote the portion of the mechanism's output that is observable to outside observers and agents $j \neq i$. A mechanism $M : \{0, 1\}^n \to \mathcal{O} \times \mathbb{R}^n_{\geq 0}$ is ϵ-jointly differentially private if for every vector of bits $\hat{b} \in \{0, 1\}^n$ for every player i, for each bit $\hat{b}'_i \in \{0, 1\}$, and for every observable set of outcomes $S \subseteq \mathcal{O} \times \mathbb{R}^{n-1}_{\geq 0}$:*

$$\Pr[M(\hat{b})_{-i} \in S] \leq \exp(\epsilon) \Pr[M(\hat{b}'_i, \hat{b}_{-i})_{-i} \in S.]$$

Note that the standard notion of differential privacy would require that a unilateral deviation by a single player i would result in only a small change in the distribution over outcomes o as well as payment vectors

π, *including* player i's own payment π_i – meaning that each player would be paid roughly the same amount, no matter what bit they report! Such a mechanism clearly cannot incentivize truthful reporting of one's bits. Joint differential privacy relaxes this constraint, and insists only that the joint distribution over the outcome o and the payments π_{-i} made to all players $j \neq i$ be insensitive to the report of player i. Crucially, π_i can depend on player i's report in an arbitrary way.

Following Nissim, Orlandi, and Tennenholtz [NOS12], we model agents as having costs for privacy $v_i(M, o, \pi_{-i})$ that can be arbitrary functions of the mechanism and the output. However, we assume that if the mechanism M is jointly differentially private, the privacy cost to an agent for all possible outcomes of the mechanism is *upper-bounded* by a linear function of the level of differential privacy: that is, if M is ϵ-jointly differentially private, $v_i(M, o, \pi_{-i}) \leq \epsilon c_i$ for all outcomes o, where c_i is agent i's personal cost coefficient. When it is clear from context, we will simply write $v(\epsilon, o, \pi_{-i})$ to emphasize that the upper bound depends on the mechanism M only via the level of joint differential privacy ϵ that M provides, and not on any other characteristic of the mechanism.

We assume that players are risk neutral, so that a player's utility for the outcome of a mechanism M that is ϵ-jointly differentially private is

$$u_i(M(\hat{b})) = \mathbb{E}[\pi_i] - \mathbb{E}[v_i(M, o, \pi_{-i})] \geq \mathbb{E}[\pi_i] - \epsilon c_i$$

We are interested in survey-like mechanisms that ask each agent to report their private bit. An agent's action choices in the mechanisms we consider are whether or not to participate, and what bit $\hat{b}_i$ to report if participating; a *strategy* for an agent i is a function σ_i mapping agent type (b_i, c_i) to an action in $\hat{b}_i \in \{0, 1, \perp\}$ (recall that $\perp$ represents non-participation). Since we are in an incomplete information setting, with a known common prior from which agents' types are drawn, we use the solution concept of a Bayes-Nash equilibrium.

Definition 2.3. *A set of strategies* $\sigma_1, \ldots, \sigma_n$ *forms a* Bayes-Nash equilibrium *if for every player i, for every realized bit* b_i *and every alternative strategy* σ_i':

$$\mathbb{E}_{(b,c) \sim \mathcal{P}}[u_i(M(\sigma(b, c)))|b_i] \geq \mathbb{E}_{(b,c) \sim \mathcal{P}}[u_i(M(\sigma_i'(b_i, c_i), \sigma_{-i}(b_{-i}, c_{-i})))|b_i].$$

Note that when we take the expectation over other player's types b_{-i}, c_{-i}, we condition only on b_i, and not on c_i: this is because of our assumption that the posterior distribution is conditionally independent of c_i given b_i.

Accuracy. The analyst is interested in computing the fraction of the population that has $b_i = 1$: we use $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} b_i$ to denote this quantity. Since agents in the population each strategically decide whether to participate and what to report as their bits, the estimate actually computed by the analyst may be different from $\hat{p}$. The analyst's goal is to design a mechanism such that its outcome closely approximates the true fraction of agents with $b_i = 1$ in the population, measured formally via its accuracy.

Definition 2.4 (Accuracy.)**.** *We say that an output estimate* $\tilde{p}$ *is* α-accurate with respect to the population if $|\hat{p} - \tilde{p}| \leq \alpha$.

We wish to design mechanisms that are α-accurate with high probability *in equilibrium*—*i.e.*, such that the estimate $\tilde{p}$ computed by the mechanism, based on the *reported* bits, from the agents who choose to participate, is within an additive error α of the true population statistic. Note that a (joint-)differentially private mechanism M cannot simply return $\tilde{p}$ as the fraction of reports $\hat{b}_i$ that are equal to 1, since it must necessarily perturb this fraction to guarantee privacy.

3 Private peer-prediction

We now address the problem of conducting sensitive surveys in settings without verifiability: how can an analyst design payments to ensure an accurate, representative outcome when she cannot verify whether privacy-sensitive agents honestly report their private information? To do this, we build on the *peer prediction* mechanism [MRZ05], designed to elicit information from strategic agents when their reports are unverifiable. The peer prediction mechanism [MRZ05], however, is not differentially private, nor does it consider agents who have privacy costs or preferences over outcomes. But modifying the peer-prediction mechanism to be differentially private also modifies the mechanism's incentive properties, since an agent's incentives for truthtelling now depend not only on whether a random reference agent will truthfully report her bit or not, but also on what fraction of the *entire* population of (strategic) potential participants enters the survey and truthfully reports their private information. In this section, we present a differentially private peer-prediction mechanism, and analyze the incentives of agents with privacy costs in this mechanism.

3.1 Peer Prediction Preliminaries

We begin by introducing some necessary background on scoring rules, which are designed for information elicitation in settings with *verifiable* outcomes.

Consider a forecasting setting, where the objective is to gather and aggregate the opinions of a set of experts about some (observable) future event. There is a large literature on the design of mechanisms—*scoring rules*—that reward an expert based on her reported prediction and the observed outcome of the event so as to incentivize the expert to truthfully reveal her true prediction about the event. Specifically, suppose there is an single expert who has some private belief about the probability of a random binary event. A *strictly proper scoring rule* provides strict incentives to the expert to truthfully report her belief, *i.e.*, the payment made by such a rule is such that the expert uniquely maximizes her expected payment by reporting her true belief.

In this paper, we will illustrate our approach to private peer-prediction via a particular strictly proper scoring rule, the well-known *Brier* scoring rule [Bri50], which makes payments for predicting a binary event as follows. Let I be the indicator random variable for the binary event to be predicted, and let q be a prediction of the probability of the event occurring. The payment for prediction q depends on the realized outcome I as

$$BasicBrier(I, q) = 2I \cdot q + 2(1 - I) \cdot (1 - q) - q^2 - (1 - q)^2,$$

and it is easy to verify that an expert who believes that I occurs with probability p will maximize her expected payment (with respect to her belief about the probability of I) by reporting $q = p$.

The following extension of the basic Brier scoring rule is central to our mechanism.

Definition 3.1 ($B(p, q)$, $B_{c,d,\rho}(p, q)$)**.** *For any p and q, we define the payment function $B(p, q)$ as follows:*

$$B(p, q) = 1 - 2(p - 2p \cdot q + q^2).$$

We also define a rescaling $B_{c,d,\rho}(p, q)$ of the payment function $B(p, q)$ as follows where $\rho > 0$:

$$B_{c,d,\rho}(p, q) = \rho(B(p - c, q - c) - d).$$

The definition of $B(p, q)$ has a simple interpretation in terms of the basic Brier scoring rule—$B(p, q)$ is exactly the *expected* payment under the basic Brier scoring rule from reporting a guess q for the probability

of an event, when the expert believes that event I occurs with probability p:

$$\mathbb{E}_{I \sim p}[BasicBrier(I, q)] = 2p \cdot q + 2(1 - p) \cdot (1 - q) - q^2 - (1 - q)^2$$
$$= 1 - 2(p - 2p \cdot q + q^2)$$
$$= B(p, q).$$

The need for the modified payment function $B_{c,d,\rho}(p, q)$ will become clear when we move to discussing private peer prediction mechanisms. The values of ρ and d simply apply a linear shift to the payoff $B(p, q)$, while the effect of c is slightly more subtle—we will use it when defining payments in the peer-prediction mechanism to symmetrize the scoring rule across observations $b_i = 0$ and $b_i = 1$.

The following easy facts about $B_{c,d,\rho}(p, q)$ will be useful in our equilibrium analysis.

Proposition 3.1. *The payment scheme $B_{c,d,\rho}(p, q)$ has the following properties.*

1. *For any p, $B_{c,d,\rho}(p, q)$ is (uniquely) maximized by reporting $q = p$.*

2. *$B_{c,d,\rho}(p, q)$ is Lipschitz-continuous in p: For any p, p',*

$$|B_{c,d,\rho}(p, q) - B_{c,d,\rho}(p', q)| \le \lambda |p - p'|, \tag{1}$$

   where $\lambda = |\rho(2 - 4(q - c))|$.

3. *Suppose the value of p is randomly drawn from a distribution with mean $\mathbb{E}[p]$. Then an agent's expected payment from reporting q is*

$$\mathbb{E}_p[B_{c,d,\rho}(p, q)] = B_{c,d,\rho}(\mathbb{E}[p], q), \tag{2}$$

   and this payment is maximized by reporting $q = \mathbb{E}[p]$.

4. *For any p, q,*
$$B_{c,d,\rho}(p, p) - B_{c,d,\rho}(p, q) = 2\rho(p - q)^2. \tag{3}$$

Proof. The first statement follows immediately by setting the derivative of $B_{c,d,\rho}(p, q)$ with respect to q to be zero since $B_{c,d,\rho}$ is strictly concave in q, the second from noting that $B_{c,d,\rho}(p, q) - B_{c,d,\rho}(p + \epsilon, q) = \rho(2 - 4(q - c))\epsilon$, and the third from noting that $B(p, q)$ is linear in p so that $\mathbb{E}_p[B(p, q)] = B(\mathbb{E}[p], q)$, and then applying the first statement. The last statement is obtained by direct substitution:

$$B_{c,d,\rho}(p, p) - B_{c,d,\rho}(p, q) = \rho(1 - 2(p - c) + 4(p - c)^2 - 2(p - c)^2 - d)$$
$$- \rho(1 - 2(p - c) + 4(p - c)(q - c) - 2(q - c)^2 - d)$$
$$= 2\rho(p - q)^2. \quad \square$$

$\square$

Peer prediction. In our setting, there is no publicly observable random event or "ground truth" whose outcome is being predicted by an "expert". Instead, we only have reports from the agents being surveyed about their private bits, albeit drawn from a common prior distribution. The key idea behind the *peer-prediction* mechanism [MRZ05] designed for such information elicitation problems with *unverifiable information* is the following: an agent's report, instead of being rewarded for its ability to predict an observable event, can

be rewarded for its success in "predicting" the outcome of the random event consisting of another agent's draw of her private bit. Note that the mechanism's only access to this private bit is via the agent's report, which necessitates introducing the notion of *equilibrium* truthful reporting: since each agent's payoff, which depends on the ability of her report to predict another agent's private bit, depends on that agent's report, we cannot just ask what a particular agent will report (as with scoring rules); rather, we must ask what reporting strategies of the population of agents constitute an equilibrium.

The vector of private bit-cost pairs $[(b_i, c_i)]$ of the n agents in our population is drawn from a (known) prior distribution $\mathcal{P}$; recall from Definition 2.1 that $\mathcal{P}_0$ (respectively $\mathcal{P}_1$) denotes the conditional distribution of the vector $b_{-i} \in \{0,1\}^{n-1}$ conditioned on $b_i = 0$ (respectively $b_i = 1$). Since these distributions $\mathcal{B}_0$ and $\mathcal{B}_1$ are known (they can be computed from the known prior $\mathcal{P}$), an agent i's draw of b_i can be used to compute a posterior probability that a random agent $j \neq i$ draws $b_j = 1$. The fact that a particular draw of b_i leads to an updated value for $\Pr(b_j = 1)$ means that one can ask agents for their bit b_i, and use it to compute their prediction of the probability that $b_j = 1$ using $\mathcal{P}$, instead of asking them to directly report their prediction of this probability.[6] Thus, if agent i reports $b_i = 0$, then the mechanism can translate that into the prediction $p_0 = \Pr[b_j = 1 | b_i = 0]$ and if agent i reports $b_i = 1$, then the mechanism can translate that into the prediction $p_1 = \Pr[b_j = 1 | b_i = 1]$.

A peer prediction mechanism based on the Brier scoring rule would then pay agent i $BasicBrier(\hat{b}_j, p_{\hat{b}_i})$ where $\hat{b}_i$, and $\hat{b}_j$ are the reports of agents i and j respectively. The first claim in Proposition 3.1, applied with $c = d = 0$ and $\rho = 1$, immediately yields that truthful reporting is a Nash equilibrium with these payments: consider an agent i with reference agent j. If j reports her bit truthfully, *i.e.*, $\hat{b}_j = b_j$, then i's belief about the probability of j's report being 1, conditional on her draw, is exactly $\Pr[b_j = 1 | b_i]$. Proposition 3.1(i) says that i maximizes her payoff in the payment rule $BasicBrier$ by reporting her true belief about this probability, namely p_{b_i}, which corresponds precisely to reporting her true bit b_i to the mechanism.

For a peer-prediction mechanism to guarantee differential privacy to agents, however, the mechanism must introduce a perturbation into the reports that are used to compute other agents' payments (recall that our notion of privacy assumes that all payments, except those made to agent i herself, are publicly visible, and therefore may reveal information about agent i. But such a perturbation will change the quantity that agent i's prediction $p_{\hat{b}_i}$ is being compared against, and correspondingly her payoffs, and possibly her incentives. We will therefore need to construct an analog of Proposition 3.1(i) that will allow us to declare that an agent will do better by truthfully reporting her bit b_i than by lying, assuming other agents also report truthfully, even when the bit he is trying to predict is not actually drawn according to the distribution $\Pr[b_j = 1 | b_i = 0]$ but only a distribution close (enough) to it.

The following proposition creates a peer-prediction mechanism that is resilient to noise, which forms the basis for designing a differentially private peer-prediction mechanism with truthful reporting in equilibrium. The first part of the proposition provides an *upper* bound on an agent's payoff from *lying* about her bit, assuming all other agents report truthfully, while the second provides a *lower* bound on the payoff from *truthful* reporting, as a function of the magnitude of the noise. Together, these two statements will allow us to lower-bound the increase in payoff from truthful reporting over lying (assuming other agents report truthfully) as a function of how much noise is added to the computation by a differentially private mechanism; we will then compare these against the improvement in privacy cost from lying to investigate when truthful reporting constitutes an equilibrium for privacy-sensitive agents, in Theorem 3.2.

---

[6]Note that eliciting b_i is not, in general, equivalent to eliciting the probability, since the set of strategies available to an agent facing a scoring rule of the form $B(p, q)$ shrinks when reporting a single bit—while in general an agent could choose to report any $q \in [0, 1]$, asking an agent to report a bit $\hat{b}_i$ restricts her to only two possible values of q, p_1 or p_0, corresponding to the reports of $\hat{b}_i = b_i$ and $\hat{b}_i = 1 - b_i$.

Proposition 3.2. *Consider a prior $\mathcal{P}$, and let $p_0, p_1 \in [0,1]$ be the posterior values $\Pr[b_j = 1|b_i]$ corresponding to this prior for $b_i = 0$ and $b_i = 1$ respectively. Let $\alpha, \beta \in \mathbb{R}$ be values such that $\beta > 0$ and $\alpha < \frac{|p_1 - p_0|}{2}$, and set the scale-and-shift parameters c, d, ρ in the modified Brier scoring rule $B_{c,d,\rho}$ to be*

$$c = (p_0 + p_1 - 1)/2,$$

$$d = \frac{1}{2} - \frac{3}{2}(p_1 - p_0)^2 + 2\alpha|p_0 - p_1|,$$

$$\rho = \frac{\beta}{2(p_1 - p_0)^2 - 4\alpha|p_0 - p_1|}.$$

Then, (i) $\rho > 0$, and (ii) the following inequalities hold for each of $b \in \{0,1\}$:

1. *For any p' with $|p' - p_b| < \alpha$, $\quad B_{c,d,\rho}(p', p_{1-b}) \leq 0$.*

2. *For any p' with $|p' - p_b| < \alpha$, $\quad \beta \leq B_{c,d,\rho}(p', p_b)$.*

3. *For any $\alpha' > 0$, p' with $|p' - p_b| < \alpha'$, $\quad \beta + 2\rho(\alpha + \alpha')|p_0 - p_1| \geq B_{c,d,\rho}(p', p_b)$.*

Proof. Note that $\rho > 0$ follows from $\beta > 0$ and $\alpha < \frac{|p_1 - p_0|}{2}$. To prove the next set of inequalities, first note that $B(1/2 + p, 1/2 + q) = B(1/2 - p, 1/2 - q)$; this can be confirmed by algebraic manipulation. Thus we have

$$B(p_0 - c, p_0 - c) = B\left(\frac{1}{2} + \frac{p_0 - p_1}{2}, \frac{1}{2} + \frac{p_0 - p_1}{2}\right)$$
$$= B\left(\frac{1}{2} + \frac{p_1 - p_0}{2}, \frac{1}{2} + \frac{p_1 - p_0}{2}\right)$$
$$= B(p_1 - c, p_1 - c),$$

and similarly

$$B(p_0 - c, p_1 - c) = B\left(\frac{1}{2} + \frac{p_0 - p_1}{2}, \frac{1}{2} + \frac{p_1 - p_0}{2}\right)$$
$$= B\left(\frac{1}{2} + \frac{p_1 - p_0}{2}, \frac{1}{2} + \frac{p_0 - p_1}{2}\right)$$
$$= B(p_1 - c, p_0 - c).$$

Next, note that for $b \in \{0,1\}$,

$$B(p_b - c, p_b - c) = B\left(\frac{1}{2} + \frac{p_b - p_{b-1}}{2}, \frac{1}{2} + \frac{p_b - p_{b-1}}{2}\right) = \frac{1}{2} + \frac{1}{2}(p_1 - p_0)^2, \tag{4}$$

where the last equality follows from the definition of $B(p, q)$. Similarly, for $b \in \{0,1\}$,

$$B(p_b - c, p_{1-b} - c) = B\left(\frac{1}{2} + \frac{p_b - p_{b-1}}{2}, \frac{1}{2} + \frac{p_{1-b} - p_b}{2}\right) = \frac{1}{2} - \frac{3}{2}(p_1 - p_0)^2. \tag{5}$$

Next, we compute the expected payoff $B_{c,d,\rho}(p_b, p_b)$ to an agent for truthfully her private bit, *assuming* all other agents report truthfully, for both possible values of the private bit $b \in \{0,1\}$:

$$
\begin{aligned}
B_{c,d,\rho}(p_b, p_b) &= \rho(B(p_b - c, p_b - c) - d) \\
&= \rho\left(\frac{1}{2} + \frac{1}{2}(p_1 - p_0)^2 - \left(\frac{1}{2} - \frac{3}{2}(p_1 - p_0)^2 + 2\alpha|p_0 - p_1|\right)\right) \\
&= \rho\left(2(p_1 - p_0)^2 - 2\alpha|p_0 - p_1|\right) \\
&= \rho(2\alpha|p_0 - p_1|) + \rho\left(2(p_1 - p_0)^2 - 4\alpha|p_0 - p_1|\right) \\
&= \rho(2\alpha|p_0 - p_1|) + \beta.
\end{aligned}
\tag{6}
$$

Similarly, the expected payoff $B_{c,d,\rho}(p_b, p_{1-b})$ to an agent for lying (*i.e.*, reporting $1 - b$ when her true bit is $b \in \{0,1\}$) *when* all other agents report truthfully, is:

$$
\begin{aligned}
B_{c,d,\rho}(p_b, p_{1-b}) &= \rho\left(B(p_b - c, p_{1-b} - c) - d\right) \\
&= \rho\left(\frac{1}{2} - \frac{3}{2}(p_1 - p_0)^2 - \left(\frac{1}{2} - \frac{3}{2}(p_1 - p_0)^2 + 2\alpha|p_0 - p_1|\right)\right) \\
&= -\rho(2\alpha|p_0 - p_1|).
\end{aligned}
\tag{7}
$$

To obtain the first statement (1), which upper-bounds the expected payoff from lying when an agent's report is compared against a noisy perturbation of other agents' true reports, we apply the Lipschitz condition of Proposition 3.1 Part 1 with $\lambda = |\rho(2-4(p_b-c))| = 2\rho|p_1-p_0|$ and $|q-p_b| \leq \alpha$ to Equation 7. Similarly, the second statement (2), which lower-bounds the expected payoff from truthtelling under noise, follows from applying the same Lipschitz condition to Equation 6. Finally, the third statement (3) follows from applying Proposition 3.1 part 1 with $\lambda = |\rho(2 - 4(p_b - c))| = 2\rho|p_1 - p_0|$ and $|q - p_b| \leq \alpha'$ to Equation 6. $\qquad\square$

3.2 A differentially-private peer prediction mechanism

We now address the question of designing incentives for a survey where privacy-sensitive individuals might incur disutility from the use of their data in this computation.

For a distribution $\mathcal{P}$ over types (b, c), recall that we have defined posterior distributions P_0 and P_1, conditioned on a draw of $b_i = 0$ and $b_i = 1$ respectively.

Our mechanism follows. Conceptually, the mechanism asks each agent to report their private bit, which they have the option of misrepresenting. From their report, the mechanism computes the posterior belief that is consistent with their report (assuming truthful reporting). It then computes the average value of all agents' reports (treating players who have opted not to participate identically as if they reported $\hat{b}_i = 0$), and perturbs this value so as to guarantee differential privacy. Finally, it pays each agent using our modified Brier scoring rule, as if they are using the computed posterior distribution to bet on the perturbed average bit value. The payments are carefully scaled to implement a Bayes Nash equilibrium in which almost all players choose to report their bit truthfully.

Mechanism $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$. Consider the following mechanism $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ for conducting a sensitive survey with prior $\mathcal{P}$, parameterized by a participation goal $1 - \alpha$, surplus payment β, and noise parameter ϵ with n agents.

1. Each participating agent i submits a report $\hat{b}_i$ of her private bit.

2. Set $\hat{b}_i = 0$ for each non-participating agent i.

3. Compute $\hat{b} = \sum_{i=1}^n \hat{b}_i$.

4. Perturb $\hat{b}$ as follows: $\bar{b} = \hat{b} + \mathrm{Lap}\left(\frac{1}{\epsilon}\right)$.

5.

$$\tilde{p} = \arg\min_{x \in [0,1]} \left| x - \frac{\bar{b}}{n} \right|$$

$$\tilde{p}_{-i} = \arg\min_{x \in [0,1]} \left| x - \frac{\bar{b} - \hat{b}_i}{n-1} \right|$$

6. Compute $p_{\hat{b}_i}$ as follows: $p_0 = \mathbb{E}_{\mathcal{P}}[\tilde{p}_{-i} | b_i = 0]$; and $p_1 = \mathbb{E}_{\mathcal{P}}[\tilde{p}_{-i} | b_i = 1]$. Here, we take the expectation over the draws of other agents' bits, assuming they are using the strategy of reporting their bits truthfully.

7. Next compute c, d, and ρ as functions of p_0, p_1, α and β as in Proposition 3.2.

8. Pay each participating agent i, π_i, based on her report $\hat{b}_i$ and $\tilde{p}_{-i}$:

$$\pi_i = B_{c,d,\rho}(\tilde{p}_{-i}, p_{\hat{b}_i}) = \rho(1 - 2((\tilde{p}_{-i} - c) - 2(\tilde{p}_{-i} - c) \cdot (p_{\hat{b}_i} - c) + (p_{\hat{b}_i} - c)^2) - d).$$

(Non-participating agents receive no payment.)

9. Output estimate $\tilde{p}$.

3.3 Analyzing $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$

In this section, we prove that $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ is an ϵ-joint differentially private mechanism supporting truthful reporting by a fraction $1 - \alpha$ of agents in equilibrium, and computes an accurate outcome.

We first show in Theorem 3.1 that $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ is ϵ-jointly differentially private, which follows from routine arguments and the following lemma.

Lemma 3.1 ("Billboard Lemma" [HHR$^+$13, RR13]). *Fix any mechanism* $\mathcal{M} : \mathcal{T}^n \to \mathcal{O}$, *for arbitrary sets* $\mathcal{T}$ *and* $\mathcal{O}$. *Fix any function* $f : \mathcal{T} \to \mathcal{O}'$. *If* $\mathcal{M}$ *is* ϵ-*differentially private, then the mechanism* $\mathcal{M}' : \mathcal{T}^n \to \mathcal{O} \times \mathcal{O}'^n$ *that computes* $o = \mathcal{M}(t)$ *and outputs* $\mathcal{M}'(t) = (o, (f(t_1, o), \ldots, f(t_n, o)))$ *is* ϵ-*jointly differentially private.*

To prove joint differential privacy, we first observe that the output $\tilde{p}$ of the mechanism observed by anyone external to the mechanism is ϵ-differentially private, and together with the output, the vector of payments π is ϵ-jointly differentially private.

Theorem 3.1. $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon} : \{0, 1, \bot\}^n \to \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}^n$ *is* ϵ-*jointly differentially private.*

Proof. Privacy follows from the fact that $\hat{b}$ has sensitivity 1, and $\bar{b}$ is computed using the Laplace mechanism of [DMNS06] with scale $1/\epsilon$. From [DMNS06], we know that the computation of $\bar{b}$ is ϵ-differentially private. This computation is followed by data independent post-processing to compute $\tilde{p}$, which cannot degrade the differential privacy guarantee. Finally, payments π_i are computed as a function $f(\hat{b}_i, \bar{b}) = B_{c,d,\rho}(\tilde{p}_{-i}, p_{\hat{b}_i})$, as allowed by Lemma 3.1. $\square$

We next prove Theorem 3.2, which addresses the question of accurate estimation of the population statistic in equilibrium by our mechanism $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$: we will show that there is a *threshold strategy equilibrium* in $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$—where all agents with cost c_i below some threshold τ participate and truthfully report their bits—that has high accuracy; this accuracy, of course, is a function of the parameters of the mechanism.

Recall that $\mathcal{P}$ is a joint distribution both over bits $b \in \{0,1\}^n$ and cost vectors $c \in \mathbb{R}^n$, and that $\mathcal{C}$ denotes the marginal distribution of $\mathcal{P}$ over cost vectors $c \in \mathbb{R}^n$. Also, $\mathcal{C}_0$ denotes the marginal distribution on cost vectors drawn from $\mathcal{P}_0$, and $\mathcal{C}_1$ the marginal distribution on cost vectors drawn from $\mathcal{P}_1$ (Definition 2.1); recall that these distributions are symmetric.

We now define a quantity $\tau_{\alpha,\delta}$, which represents a cost threshold that satisfies two conditions. First, it represents a threshold such that with high probability $1 - \delta$ (with respect to the prior $\mathcal{P}$), at least a $1 - \alpha$ fraction of individuals in the population have costs $c_i \leq \tau_{\alpha,\delta}$. Second, it is *also* a threshold such that for every player i, conditioned on seeing either the bit $b_i = 1$ or $b_i = 0$, the probability that a random other agent $j \neq i$ has cost $c_j \leq \tau_{\alpha,\delta}$ is at least $1 - \alpha$.

Definition 3.2 (Threshold $\tau_{\alpha,\delta}$.). *Fix a prior $\mathcal{P}$ with a corresponding marginal cost distribution $\mathcal{C}$, and let*

$$\tau_{\alpha,\delta}^1 = \inf_\tau \left(\Pr_{c \sim \mathcal{C}} [|\{i : c_i \leq \tau\}| \geq (1-\alpha)n] \geq 1 - \delta \right),$$

$$\tau_\alpha^2 = \inf_\tau \left(\min \left(\Pr_{c \sim \mathcal{C}_0, j \neq i} [c_j \leq \tau], \Pr_{c \sim \mathcal{C}_1, j \neq i} [c_j \leq \tau] \right) \geq 1 - \alpha \right).$$

We define $\tau_{\alpha,\delta}$ as the larger of these two thresholds:

$$\tau_{\alpha,\delta}^1 = \max(\tau_{\alpha,\delta}^1, \tau_\alpha^2).$$

We present the following theorem.

Theorem 3.2. *Let $\mathcal{P}$ be a symmetric prior over types (b,c) satisfying the condition that for every i, the posterior distribution $\mathcal{P}_{b_i}$ on $(b,c)_{-i}$ given b_i is conditionally independent of c_i. Fix a participation goal $1 - \alpha$ such that $\alpha < \frac{|p_1 - p_0|}{2}$, a privacy parameter ϵ, and the desired confidence δ on the accuracy guarantee. If the parameter β in mechanism $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ is chosen to be $\beta = \epsilon \tau_{\alpha,\delta/2}$, then:*

1. *The mechanism $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ has a symmetric Bayes Nash equilibrium consisting of threshold strategies $\sigma_{\tau_{\alpha,\delta/2}}$, where $\sigma_{\tau_{\alpha,\delta/2}}(b_i, c_i) = b_i$ whenever $c_i \leq \tau_{\alpha,\delta/2}$, i.e., all agents with cost smaller than $\tau_{\alpha,\delta/2}$ participate and truthfully report their private bit ($\sigma_{\tau_{\alpha,\delta/2}}$ can be arbitrary for agents with $c_i > \tau_{\alpha,\delta/2}$.)*

2. *In the equilibrium $\sigma_{\tau_{\alpha,\delta/2}}$, the estimate $\tilde{p}$ computed by $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ is α'-accurate for $\alpha' = \left(\frac{\ln(\delta/2)}{\epsilon n} + \alpha \right)$ with probability at least $1 - \delta$, where the probability is over the draw of types from $\mathcal{P}$.*

Remark 3.1. *If we want a mechanism which is 2α accurate, it suffices to take $\epsilon = \frac{\ln(1/\delta)}{\alpha n}$, in which case our β parameter is set as:*

$$\beta = \frac{\ln(1/\delta) \cdot \tau_{\alpha,\delta/2}}{\alpha n}$$

Proof. We first show (1), that the threshold strategies $\sigma_{\tau_{\alpha,\delta/2}}$ form a Bayes Nash equilibrium.

An agent i's payoff if she participates in $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ is the difference between her payment π_i and her (dis)utility from her privacy loss, $v_i(\epsilon, o)$. To analyze agents' decisions in $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$, note first that every agent's payoff depends on the decisions of all other agents in *two* ways—the payment $\pi_i = B_{c,d,\rho}(\tilde{p}_{-i}, p_{\hat{b}_i})$

14

uses the reports $\hat{b}_{-i}$ of all other agents in computing $\tilde{p}_{-i}$, and the cost $v_i(\epsilon, o)$ depends on o which in turn is computed using the reports of all other agents.

Suppose all the agents other than agent i are all following a symmetric *threshold* strategy $\sigma_{j,\tau_{\alpha,\delta/2}}$ that results in truthful reporting of $\sigma_{j,\tau_{\alpha,\delta/2}}(b_j, c_j) = b_j$ whenever $c_j \leq \tau_{\alpha,\delta/2}$, and can result in arbitrary behavior otherwise.

We can now consider whether the same threshold strategy $\sigma_{i,\tau_{\alpha,\delta/2}}$ is a best response for player i: if it is, then we will have shown that $\sigma_{\tau_{\alpha,\delta/2}}$ forms a symmetric Bayes Nash equilibrium. Assume that $c_i < \tau_{\alpha,\delta/2}$ as otherwise there is nothing to show, since the strategy does not specify agents' actions if their cost exceeds the threshold.

Consider agent i's incentives, given reports $\hat{b}_{-i}$ from the remaining participants in the mechanism. Agent i chooses her strategy from three possible options: $\{b_i, 1-b_i, \perp\}$—truth-telling, lying, or not participating—by evaluating her expected utility given the strategy choices of the remaining agents, where the expectation is over the remaining agents' draws of bit, cost pairs $(b, c)_{-i}$ from $\mathcal{P}$, as well as the random coin tosses in $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$. Agent i has an incentive to report the truth in $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ if her expected payoff from reporting $\hat{b}_i = b_i$ is at least as large as the payoff from either reporting $\hat{b}_i = 1 - b_i$ or not participating.

Let S_{-i} denote the individuals, other than i, who truthfully participate when using strategy $\sigma_{\tau_{\alpha,\delta/2}}$. Note that by the definition of $\sigma_{\tau_{\alpha,\delta/2}}$ we have $\mathbb{E}[|S_{-i}|] \geq (1-\alpha)(n-1)$, since each individual truthfully reports with probability (over his type distribution) at least $1-\alpha$. Thus for $b_i \in \{0,1\}$ we have

$$|\mathbb{E}[\tilde{p}_{-i}|b_i] - \mathbb{E}[p_{b_i}|b_i]| \leq \alpha.$$

We first compute the payoff for truth-telling, $\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-1}, p_{b_i})|b_i] - \mathbb{E}_{o\sim M(b_i,b_{-i})}[v_i(\epsilon, o)]$ and show that it is positive. Above we saw that $|\mathbb{E}[\tilde{p}_{-i}|b_i] - \mathbb{E}[p_{b_i}|b_i]| \leq \alpha$. Thus, by our choice of c, d, ρ and Proposition 3.2 we have that

$$\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_i})|b_i] \geq \beta = \epsilon\tau_{\alpha,\delta/2}.$$

However, we also know that for each player i, $\mathbb{E}_{o\sim M(b_i,b_{-i})}[v_i(\epsilon, o)] \leq \epsilon c_i$ by the ϵ-joint differential-privacy of the mechanism. By assumption $\tau_{\alpha,\delta/2} > c_i$, and so it follows that truth-telling has a positive expected payoff for agent i.

We now show that the expected payoff for lying, $\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{1-b_i})|b_i] - \mathbb{E}_{o\sim M(1-b_i,b_{-i})}[v_i(\epsilon, o)]$ is always non-positive. Above we saw that $|\mathbb{E}[\tilde{p}_{-i}|b_i] - \mathbb{E}[p_{b_i}|b_i]| \leq \alpha$. Thus, by our choice of c, d, ρ and Proposition 3.2 we have that $\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{1-b_i})|b_i] \leq 0$. The second term is non-positive, so it follows that the payoff for lying is also non-positive.

Finally, we look at the payoff for non-participation, which is simply $-\mathbb{E}_{o\sim M(\perp,\hat{b}_{-i})}[v_i(\epsilon, o)]$ and non-positive by definition.

Because the utility agent i receives from for truth-telling is nonnegative, but her utility from lying or non-participation is non-positive, she is always at least as well off truth-telling.

It remains to show (2), that with probability at least $1 - \delta$ the mechanism is α'-accurate for $\alpha' = \left(\frac{\ln(\delta/2)}{\epsilon n} + \alpha\right)$-when all players play according to equilibrium $\sigma_{\tau_{\alpha,\delta/2}}$. To show this, we note that error comes from two sources: first, non-truthful behavior. However, by the definition of $\sigma_{\tau_{\alpha,\delta/2}}$, we know that with probability at least $1-\delta/2$, all but $\alpha n/2$ people truth-tell in equilibrium: that is, except with probability $1 - \delta/2$, we have $|\sum_{i=1}^n \hat{b}_i - \sum_{i=1}^n b_i| \leq \frac{\alpha n}{2}$. The second source of error is the Laplace noise added to $\hat{b}$. We have that $\Pr[|\text{Lap}(1/\epsilon)| \geq t/\epsilon] = \exp(-t)$. Hence, $\Pr[|\hat{b} - \bar{b}| \geq \ln(\delta/2)/\epsilon] \leq \delta/2$. Combining these two bounds gives the claim. $\qquad\square$

Remark 3.2. *We remark that the above analysis easily extends to the case where privacy-sensitive individuals might incur disutility from the outcome o of the analyst's computation, as well as from the use of their data in this computation. In this case, assume that an agent's cost c_i bounds both the cost of his privacy and the difference in the agent's utility between any two outcomes. Because the mechanism is ϵ private, each agent has at most ϵ influence on its outcome. Hence the cost in the outcome quality for changing her bit (or not participating) is at most $c_i\epsilon$: $\mathbb{E}_{o \sim M(b_i, b_{-i})}[v_i(\epsilon, o)] - \mathbb{E}_{o \sim M(b'_i, b_{-i})}[v_i(\epsilon, o)] \leq \epsilon(\max_{o, o'} v_i(\epsilon, o) - v_i(\epsilon, o')) \leq c_i\epsilon$.*

3.4 An alternative model of privacy costs

In the previous sections, we have focused on a conservative model of agents' costs for particular outcomes and privacy which assumes only that agent costs are upper bounded by a linear function of the differential privacy guarantee. This model was proposed by [NOS12], and is the weakest assumption seen in the differential privacy and game theory literature. In this section, we briefly consider a less-conservative, but still well-motivated, assumption on privacy cost functions proposed by Chen et al. [CCK+13]. Under this stronger assumption, we are able to show not only that it is possible to truthfully elicit agents' private bits and compute on them accurately, but that the privacy cost of doing so tends to 0 as n grows large. In fact, because non-private peer-prediction payments can be arbitrarily rescaled, the Chen et al. model implies that by gathering more participants, a surveyor could drive the entire cost of running our private peer prediction mechanism to zero.[7]

The following assumption is designed to capture the intuition that an agent should experience low privacy cost for outcomes that would induce only a small change in a Bayesian adversary's beliefs about her type (see [CCK+13] for a thorough motivation).

Assumption 3.1 ([CCK+13] Privacy Cost Assumption)**.** *We assume that for any mechanism $M : \{0,1\}^n \to \mathcal{O} \times \mathbb{R}^n_{\geq 0}$, $\forall \hat{b} \in \{0,1\}^n$, for all players i, $o \in \mathcal{O}$, $\pi_{-i} \in \mathbb{R}^{n-1}_{\geq 0}$:*

$$\left| v_i(M, (o, \pi_{-i}), \hat{b}_i, \hat{b}_{-i}) \right| \leq c_i \ln \left(\max_{b'_i, b''_i \in \{0,1\}} \frac{\Pr[M(b'_i, \hat{b}_{-i}) = (o, \pi_{-i})]}{\Pr[M(b''_i, \hat{b}_{-i}) = (o, \pi_{-i})]} \right).$$

where the probabilities are taken over the random choices of M.[8]

They show a useful lemma bounding the expected privacy loss of agents participating in a differentially private computation who experience privacy costs according to this assumption:

Lemma 3.2 ([CCK+13] Composition Lemma)**.** *In settings that obey the above cost assumption, and for mechanisms M that are ϵ-differentially private for $\epsilon \leq 1$, then for a player i with bit b_i, $\forall b_{-i} \in \{0,1\}^{n-1}$, $\forall b'_i \in \{0,1\}$,*

$$\mathbb{E}[v_i(M, M(b), b_i, b_{-i})] - \mathbb{E}[v_i(M, M(b'_i, b_{-i}), b_i, b_{-i})] \leq 2c_i\epsilon(e^\epsilon - 1) \leq 4c_i\epsilon^2$$

With the above assumption, we can achieve an accurate Bayes-Nash equilibrium while scaling payments down with ϵ at a rate of ϵ^2, rather than linearly. As we will see, this will allow us to drive our total costs down to zero.

---

[7]This model does not incorporate a minimum base cost of approaching each potential survey participant.

[8]The model proposed in [CCK+13] is somewhat more general, and allows the costs to be bounded by arbitrary functions of the log probability ratio, not just linear functions. We adopt this linear model here for simplicity, so that we do not need to modify our model of having a distribution of costs c_i.

We first show an analogue of Theorem 3.2. The only difference is that with our new stronger assumption on agent privacy costs, we can obtain a Bayes Nash equilibrium while taking the β parameter to be substantially smaller than before.

Theorem 3.3. *Suppose agents' privacy costs satisfy Assumption 3.1. Let $\mathcal{P}$ be a symmetric prior over types (b, c) satisfying the condition that for every i, the posterior distribution $\mathcal{P}_{b_i}$ on $(b, c)_{-i}$ given b_i is conditionally independent of c_i. Fix a participation goal $1 - \alpha$ such that $\alpha < \frac{|p_1 - p_0|}{2}$, a privacy parameter ϵ, and the desired confidence δ on the accuracy guarantee. If the parameter β of the mechanism is chosen to be $\beta = 4\epsilon^2 \tau_{\alpha,\delta/2}$, then:*

1. *The mechanism $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ has a symmetric Bayes Nash equilibrium consisting of threshold strategies $\sigma_{\tau_{\alpha,\delta/2}}$, where $\sigma_{\tau_{\alpha,\delta/2}}(b_i, c_i) = b_i$ whenever $c_i \leq \tau_{\alpha,\delta/2}$, i.e., all agents with cost smaller than $\tau_{\alpha,\delta/2}$ participate and truthfully report their private bit ($\sigma_{\tau_{\alpha,\delta/2}}$ can be arbitrary for agents with $c_i > \tau_{\alpha,\delta/2}$.)*

2. *In the equilibrium $\sigma_{\tau_{\alpha,\delta/2}}$, the estimate $\tilde{p}$ computed by $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ is α'-accurate for $\alpha' = \left(\frac{\ln(\delta/2)}{\epsilon n} + \alpha \right)$ with probability at least $1 - \delta$, where the probability is over the draw of types from $\mathcal{P}$.*

Remark 3.3. *Note that this theorem differs from our previous theorem in how it sets β. Since we can again take $\epsilon = \frac{\ln(1/\delta)}{\alpha n}$, we have*

$$\beta = \frac{4\ln(1/\delta)^2 \cdot \tau_{\alpha,\delta/2}}{\alpha^2 n^2}$$

Proof. We first show (1), that the threshold strategies $\sigma_{\tau_{\alpha,\delta/2}}$ form a Bayes Nash equilibrium.

Suppose the agents $j \neq i$ are all following a symmetric *threshold* strategy $\sigma_{j,\tau_{\alpha,\delta/2}}$ that results in truthful reporting of $\sigma_{j,\tau_{\alpha,\delta/2}}(b_j, c_j) = b_j$ whenever $c_j \leq \tau_{\alpha,\delta/2}$, and can result in arbitrary behavior otherwise.

We can now consider whether the same threshold strategy $\sigma_{i,\tau_{\alpha,\delta/2}}$ is a best response for player i: if it is, then we will have shown that $\sigma_{\tau_{\alpha,\delta/2}}$ forms a symmetric Bayes Nash equilibrium. Assume that $c_i < \tau_{\alpha,\delta/2}$ as otherwise there is nothing to show.

An agent's i's payoff if she participates in $\mathcal{M}_{\mathcal{P},n,\alpha,\beta,\epsilon}$ is the difference between her payment π_i and her (dis)utility from her privacy loss, $v_i(\epsilon, o)$. We will first show that an agent's utility for participating and truth-telling is higher than her payment for participating and misrepresenting her bit.

An agent i has a higher payoff for truth-telling than for lying if:

$$\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_i})|b_i] - \mathbb{E}[v_i(M, M(\hat{b}), \hat{b}_i, \hat{b}_{-i})] \geq \mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_{1-i}})|b_i] - \mathbb{E}[v_i(M, M(\hat{b}), 1 - \hat{b}_i, \hat{b}_{-i})]$$

Or equivalently,

$$\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_i})|b_i] - \mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_{1-i}})|b_i] \geq \mathbb{E}[v_i(M, M(\hat{b}), \hat{b}_i, \hat{b}_{-i})] - \mathbb{E}[v_i(M, M(\hat{b}), 1 - \hat{b}_i, \hat{b}_{-i})].$$

By Assumption 3.1 and Lemma 3.2, the right hand side of the above inequality is at most $4c_i\epsilon^2$, since the mechanism is ϵ-differentially private. It remains to bound the left-hand side.

Let S_{-i} denote the individuals, other than i, who truthfully participate when using strategy σ_τ. Note that by the definition of $\sigma_{\tau_{\alpha,\delta/2}}$ we have $\mathbb{E}[|S_{-i}|] \geq (1 - \alpha)(n - 1)$, since each individual truthfully reports with probability (over his type distribution) at least $1 - \alpha$. Thus for $b_i \in \{0, 1\}$,

$$|\mathbb{E}[\tilde{p}_{-i}|b_i] - \mathbb{E}[p_{b_i}|b_i]| \leq \alpha.$$

Above we saw that $|\mathbb{E}[\tilde{p}_{-i}|b_i] - \mathbb{E}[p_{b_i}|b_i]| \leq \alpha$. Thus, by our choice of c, d, ρ and Proposition 3.2,

$$\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_i})|b_i] \geq \beta = 4\epsilon^2 \tau_{\alpha,\delta/2}.$$

By assumption $\tau_{\alpha,\delta/2} \geq c_i$, and so if we can show that the expected payment that a player receives for misreporting his bit is non-positive, we will be done.

We now show that the expected payoff for lying, $\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{1-b_i})|b_i]$ is always non-positive. Above we saw that $|\mathbb{E}[\tilde{p}_{-i}|b_i] - \mathbb{E}[p_{b_i}|b_i]| \leq \alpha$. Thus, by our choice of c, d, ρ and Proposition 3.2 we have that $\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{1-b_i})|b_i] \leq 0$. The second term is non-positive, so it follows that the payoff for lying is also non-positive.

We similarly must show that the utility associated with participation and truth-telling is greater than the utility of non-participation. That is:

$$\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_i})|b_i] - \mathbb{E}[v_i(M, M(\hat{b}), \hat{b}_i, \hat{b}_{-i})] \geq -\mathbb{E}[v_i(M, M(\hat{b}), \perp, \hat{b}_{-i})]$$

This statement follows from analysis analogous to showing that truth-telling outperforms lying, since non-participation induces the same privacy cost as reporting $\hat{b}_i = 0$, but results in no payment.

It remains to show (2), that with probability at least $1 - \delta$ the mechanism is α'-accurate for $\alpha' = \left(\frac{\ln(\delta/2)}{\epsilon n} + \alpha\right)$-when all players play according to equilibrium $\sigma_{\tau_{\alpha,\delta/2}}$. To show this, we note that error comes from two sources: first, non-truthful behavior. However, by the definition of $\sigma_{\tau_{\alpha,\delta/2}}$, we know that with probability at least $1 - \delta/2$, all but $\alpha n/2$ people truth-tell in equilibrium: that is, except with probability $1 - \delta/2$, we have $|\sum_{i=1}^n \hat{b}_i - \sum_{i=1}^n b_i| \leq \frac{\alpha n}{2}$. The second source of error is the Laplace noise added to $\hat{b}$. We have that $\Pr[|\mathrm{Lap}(1/\epsilon)| \geq t/\epsilon] = \exp(-t)$. Hence, $\Pr[|\hat{b} - \bar{b}| \geq \ln(\delta/2)/\epsilon] \leq \delta/2$. Combining these two bounds gives the claim. $\square$

Finally, we show that the total payment that the surveyor need produce tends to zero as the population size n grows large. In particular, this means that the marginal cost of preserving privacy while doing peer prediction tends to zero as n grows large.

Theorem 3.4. *Fix any parameters α and δ. In the symmetric Bayes Nash equilibrium defined by threshold strategies $\sigma_{\tau_{\alpha,\delta/2}}$, the total expected cost incurred by the surveyor:*

$$\sum_{i=1}^n \mathbb{E}[\pi_i] \leq \frac{4\ln(1/\delta)^2 \cdot \tau_{\alpha,\delta/2}}{\alpha^2 n} \cdot \left(1 + 4\frac{\alpha}{2(p_1 - p_0) - 4\alpha}\right) = O\left(\frac{1}{n}\right)$$

Proof. Recall that in equilibrium, players maximize their payoff by truthtelling. Therefore:

$$\sum_{i=1}^n \mathbb{E}[\pi_i] \leq \sum_{i=1}^n \mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_i})|b_i]$$

However, recall that we have shown that $|\mathbb{E}[\tilde{p}_{-i}|b_i] - \mathbb{E}[p_{b_i}|b_i]| \leq \alpha$. Therefore, we can invoke proposition 3.2 (3) to conclude:

$$
\begin{aligned}
\mathbb{E}[B_{c,d,\rho}(\tilde{p}_{-i}, p_{b_i})|b_i] &\leq \beta + 4\rho\alpha|p_0 - p_1| \\
&= \beta \cdot \left(1 + 4\frac{\alpha|p_0 - p_1|}{2(p_1 - p_0)^2 - 4\alpha|p_0 - p_1|}\right) \\
&= \frac{4\ln(1/\delta)^2 \cdot \tau_{\alpha,\delta/2}}{\alpha^2 n^2} \cdot \left(1 + 4\frac{\alpha}{2(p_1 - p_0) - 4\alpha}\right)
\end{aligned}
$$

Therefore:

$$\sum_{i=1}^n \mathbb{E}[\pi_i] \leq \frac{4\ln(1/\delta)^2 \cdot \tau_{\alpha,\delta/2}}{\alpha^2 n} \cdot \left(1 + 4\frac{\alpha}{2(p_1 - p_0) - 4\alpha}\right) = O\left(\frac{1}{n}\right). \quad \square$$

18

□

4 Conclusion

In this paper we have shown how to *accurately* conduct a survey when agents have costs associated with their privacy loss that might result from the survey outcome, even when the surveyor has no ability to verify agents' private data. This result holds under even an extremely mild assumption on agents' privacy costs—that they are simply *upper bounded* as a function of the privacy parameter ϵ when the mechanism satisfies ϵ-differential privacy. Under a stronger, but still well-motivated assumption on privacy costs, we have further shown that remarkably, the *cost* of conducting such a survey (and, in particular, the marginal cost of compensating for privacy losses) can be driven to zero simply by increasing the number of participants.

However, compared to past literature on the sensitive surveyors problem, our results come at a price. Because we cannot verify agents' bits, we inherit from the peer-prediction literature that truth-telling results in just one of possibly many Bayes Nash equilibria of our mechanism. Although the truth-telling equilibrium might be considered "focal", and hence likely to occur given a lack of coordination by the agents, in the presence of coordination, we might be worried that agents can collude and coordinate on a different equilibrium, which might result in lower privacy costs for them. The problem of eliminating these bad equilibria is an exciting direction for future work.

References

[Bri50]   Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[CCK+13] Yiling Chen, Stephen Chong, Ian A Kash, Tal Moran, and Salil Vadhan. Truthful mechanisms for agents that value privacy. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 215–232. ACM, 2013.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*, pages 265–284, 2006.

[FL12]   Lisa Fleischer and Yu-Han Lyu. Approximately optimal auctions for selling privacy when costs are correlated with data. In *ACM Electronic Commerce (EC)*, pages 568–585, 2012.

[GL13]   Arpita Ghosh and Katrina Ligett. Privacy and coordination: computing on databases with endogenous participation. In *ACM Conference on Electronic Commerce*, pages 543–560, 2013.

[GR13]   Arpita Ghosh and Aaron Roth. Selling privacy at auction. *Games and Economic Behavior*, 2013. Preliminary Version appeared un the Proceedings of the Twelfth ACM Conference on Electronic Commerce (EC 2011).

[GRP09]   Sharad Goel, Daniel M. Reeves, and David M. Pennock. Collective revelation: A mechanism for self-verified, weighted, and truthful predictions. In *Proceedings of the 10th ACM conference on Electronic commerce (EC 2009)*, 2009.

[HHR+13] Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *arXiv preprint arXiv:1311.2828*, 2013.

[JF06]   Radu Jurca and Boi Faltings. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce (EC 2006)*, 2006.

[JF07]   Radu Jurca and Boi Faltings. Robust incentive-compatible feedback payments. In *Trust, Reputation and Security: Theories and Practice*, volume 4452, pages 204–218. Springer-Verlag, 2007.

[JF08]   Radu Jurca and Boi Faltings. Incentives for expressing opinions in online polls. In *Proceedings of the 9th ACM conference on Electronic commerce (EC 2008)*, 2008.

[JF09]   Radu Jurca and Boi Faltings. Mechanisms for making crowds truthful. *J. Artif. Int. Res.*, 34(1), March 2009.

[KPRU14] Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 403–410. ACM, 2014.

[LR12]   Katrina Ligett and Aaron Roth. Take it or leave it: Running a survey when privacy comes at a cost. In *Internet and Network Economics*, pages 378–391. Springer, 2012.

[LS08]   N. Lambert and Y. Shoham. Truthful surveys. *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE 2008)*, 2008.

[MRZ05]   N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, pages 1359–1373, 2005.

[NOS12]   Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 774–789. ACM, 2012.

[NVX14]   Kobbi Nissim, Salil Vadhan, and David Xiao. Redrawing the boundaries on purchasing data from privacy-sensitive individuals. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 411–422. ACM, 2014.

[PR13]   Mallesh Pai and Aaron Roth. Privacy and mechanism design. *Sigecom Exchanges*, pages 8–29, 2013.

[Pre04]   D. Prelec. A Bayesian Truth Serum for subjective data. *Science*, 306(5695):462–466, 2004.

[PRGJ11] A. Papakonstantinou, A. Rogers, E.H. Gerding, and N.R. Jennings. Mechanism design for the truthful elicitation of costly probabilistic estimates in distributed information systems. *Artificial Intelligence*, 175(2):648–672, 2011.

[RR13]   Ryan Rogers and Aaron Roth. Asymptotically truthful equilibrium selection in large congestion games. *arXiv preprint arXiv:1311.2625*, 2013.

[RS12]   Aaron Roth and Grant Schoenebeck. Conducting truthful surveys, cheaply. In *ACM Electronic Commerce (EC)*, pages 826–843, 2012.

[WP11]   Jens Witkowski and David C. Parkes. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*, 2011.

[WP12a]   J. Witkowski and D. Parkes. A robust Bayesian Truth Serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*, 2012.

[WP12b]   Jens Witkowski and David C. Parkes. Peer Prediction without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC 2012)*, 2012.