Truthful Mechanisms for Agents that Value Privacy

YILING CHEN, Harvard University
STEPHEN CHONG, Harvard University
IAN A. KASH, Microsoft Research
TAL MORAN, Efi Arazi School of Computer Science, IDC Herzliya
SALIL VADHAN, Harvard University

1. INTRODUCTION

In this paper, we examine the interaction between mechanism design and differential privacy. In particular, we explicitly model privacy in players' utility functions and design truthful mechanisms with respect to it. Our work is motivated by considerations in both fields.

In mechanism design, it has long been recognized that players may not behave as predicted due to traditional incentives analysis out of concerns for privacy: in addition to having preferences about the outcome of a mechanism (e.g., who wins an auction, or where a hospital is located), they may also be concerned about what others learn about their private information (e.g., how much they value the auctioned good, or whether they have some medical condition that makes them care more about the hospital's location). The latter concerns are not modelled in most works on mechanism design, and it is natural to try to bring the new models and techniques of differential privacy to bear on them.

Differential privacy [Dwork et al. 2006] is a notion developed to capture privacy when performing statistical analyses of databases. Informally, a randomized algorithm is differentially private if changing a single individual's data does not "substantially" change the output distribution of the algorithm. Thus, differential privacy is not an absolute notion, but rather a quantitative one that needs to be weighed against other objectives. Indeed, differentially private algorithms typically offer a tradeoff between the level of privacy offered to individuals in a database and the accuracy of statistics computed on the database, which we can think of as a "global" objective to be optimized. However, it is also of interest to consider how privacy should be weighed against the objectives of the individuals themselves. Mechanism design provides a natural setting in which to consider such tradeoffs. Attempting to model and reason about privacy in the context of mechanism design seems likely to lead to an improved understanding about the meaning and value of privacy.

1.1. Previous Work

The first work bringing together differential privacy and mechanism design was by McSherry and Talwar [2007]. They showed how to use differential privacy as a *tool* for mechanism design. By

Work begun when all the authors were at the Harvard Center for Computation and Society. Supported in part by a gift from Google, Inc.; NSF grants CCF-0915016, CNS-1237235, and 1054172; and a grant from the Sloan Foundation. Author's addresses:

Y. Chen, S. Chong, and S. Vadhan, Center for Research on Computation and Society and School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA;
I. A. Kash, Microsoft Research Cambridge, 21 Station Road, Cambridge CB1 2FB, UK;
T. Moran, Efi Arazi School of Computer Science, IDC Herzliya.

DOI:http://dx.doi.org/10.1145/0000000.0000000

definition, differentially private algorithms are insensitive to individuals' inputs; a change in a single individual's input has only a small effect on the output distribution of the algorithm. Thus, if a mechanism is differentially private (and players have bounded utility functions), it immediately follows that the mechanism is *approximately truthful*. That is, reporting untruthfully can only provide a small gain in a player's utility. With this observation, McSherry and Talwar showed how tools from differential privacy allow construction of approximately truthful mechanisms for many problems, including ones where exact truthfulness is impossible.

However, as pointed out by Nissim, Smorodinsky, and Tennenholz [2010], the approximate truthfulness achieved by McSherry and Talwar [2007] may not be a satisfactory solution concept. While differential privacy can guarantee that a player will gain arbitrarily little by lying, it also makes the potential gain from telling the truth equally small. Thus players may choose to lie in order to protect their privacy. Even worse, as shown by an example in Nissim et al. [2010], in some cases misreporting is a dominant strategy of the game. Thus, it is difficult to predict the outcome and "global" objectives such as social welfare of differentially private mechanisms. Motivated by this, Nissim et al. [2010] show how to modify some of the mechanisms of McSherry and Talwar [2007] to provide exact truthfulness. To do so, however, they sacrifice differential privacy and/or augment the standard mechanism design framework so that the mechanism can restrict the way in which players can take advantage of the selected outcome.

A recent paper by Xiao [2011; 2013] shows how to remedy this deficiency and construct mechanisms that simultaneously achieve exact truthfulness and differential privacy. Xiao's paper also points out that even this combination may not be sufficient for getting players that value privacy to report truthfully. Indeed, exact truthfulness only means that a player *weakly* prefers to tell the truth. Lying might not reduce the player's utility at all (and differential privacy implies that it can only reduce the player's utility by at most a small amount). On the other hand, differential privacy does not guarantee "perfect" privacy protection, so it is possible that a player's concern for privacy may still outweigh the small or zero benefit from being truthful.

To address this, Xiao advocated incorporating privacy directly into the players' utility functions, and seeking mechanisms that are truthful when taking the combined utilities into account. He proposed to measure privacy cost as the the mutual information between a player's type (assumed to come from some prior distribution) and the outcome of the mechanism.[1] Using this measure, he showed that his mechanism does not remain truthful when incorporating privacy into the utility functions, and left as an open problem to construct mechanisms that do. The difficulty is that we can only incentivize truthfulness by giving players an influence on the outcome, but such an influence also leads to privacy costs, which may incentivize lying. Reconciling this tension requires finding models and mechanisms where the privacy costs can be shown to be quantitatively dominated by the outcome-utility benefits of truthfulness.

1.2. Our Contributions

In this paper, we propose a new, more general way of modelling privacy in players' utility functions. Unlike Xiao's mutual information measure, our model does not require assuming a prior on players' types, and is instead a pointwise model: we simply assume that if an outcome o has the property that any report of player i would have led to o with approximately the same probability, then o has small privacy cost to player i. One motivation for this assumption is that such an outcome o will induce only a small change in a Bayesian adversary's beliefs about player i (conditioned on the other players' reports). (This is inspired by a Bayesian interpretation of differential privacy, due to Dwork and McSherry and described in [Kasiviswanathan and Smith 2008].) While Xiao's mutual information measure is not strictly a special case of our model, we show (in the appendix) that truthfulness with respect to our modelling implies truthfulness with respect to Xiao's.

[1] The mutual information measure is from the original version of Xiao's paper [Xiao 2011]. Subsequent to our work, he has revised his model to use a different, prior-free measure of privacy [Xiao 2013].

With this modelling, we show that it is possible to construct mechanisms where the privacy costs are dominated by the outcome-utility benefits of truthfulness. Specifically, we give three mechanisms that are truthful with respect to our model of privacy: one for an election between two candidates, one for a discrete version of the facility location problem, and one for general social choice problems with discrete utilities (via a VCG-like mechanism). As the number n of players increases, the social welfare achieved by our mechanisms approaches optimal (as a fraction of n).

To illustrate our ideas, consider an election between two candidates, where each player has a noticeable preference for one candidate over the other (but this preference could potentially be outweighed by privacy concerns). To carry out such an election in a differentially private way, we can use a noisy majority vote. That is, we take a majority rule after adding random noise of magnitude $O(1/\epsilon)$, where ϵ is the "differential privacy parameter" that bounds the influence of each player's vote on the outcome. Now, as we reduce ϵ, we also reduce the privacy costs experienced by each player. However, simply bounding the *overall* privacy cost by a function of ϵ does not suffice to argue truthfulness of the election mechanism. Indeed, consider a highly skewed election, where $2/3$ of the voters prefer the first candidate. Then the chance that any single player's vote affects the outcome is exponentially small in ϵn, where n is the number of voters. If the player's overall privacy cost can be as large as ϵ, then we would need to set ϵ very small (namely $\tilde{O}(1/n)$) and hence add a very large amount of noise (sacrificing social welfare) in order to achieve truthfulness. However, we argue that the overall privacy cost of being truthful in a skewed election should also be exponentially small. Indeed, for most settings of the random noise, a player cannot affect the outcome by misreporting, so the privacy cost experienced by the player should be the same regardless of whether the player reported truthfully or not. For the remaining settings of the random noise, in which the player can affect the outcome, reporting truthfully provides the player with a noticeable gain in outcome utility, whereas the privacy costs are bounded by a function of ϵ and hence can be made arbitrarily small. Put differently, the *expected* change in privacy cost experienced by a player that changes her report is bounded by ϵ times the probability that the outcome changes (which is always at most $O(\epsilon)$, but can be much smaller, as in the case of a skewed election). This analysis crucially uses the fact that we adopt a *per-outcome* model of privacy, and can be viewed as a generalization of the result of [Dwork et al. 2010] showing that every ϵ-differentially private mechanism actually has expected privacy loss $O(\epsilon^2)$.

Our mechanism for discrete facility location is inspired by Xiao's mechanism [Xiao 2013] (which he shows to be differentially private and truthful if we don't incorporate privacy into the utility functions), but with some variation in order to be apply an analysis like the above. For both the election and facility location mechanisms, an analysis like the one sketched above establishes *universal truthfulness*—truthfulness for every choice of the mechanism's random coins. For our VCG-like mechanism for general social choice problems, we need to work a bit harder to also ensure that the payments requested do not compromise privacy, and this leads us to only achieve truthfulness in expectation.

Unlike previous works, we do not treat differential privacy as an end in itself but rather as a means to incentivize truthfulness from agents that value privacy. Thus, we do not necessarily need to set the differential privacy parameter ϵ to be very small (corresponding to very high privacy, but a larger price in social welfare); we only need to set it small enough so that the privacy costs are outweighed by the agents' preferences for outcomes. Specifically, our analysis shows that as we decrease ϵ, agents' ability to affect the outcome falls, but their expected privacy cost falls even faster. Thus, it is natural to conclude (as we do) that there is some value of ϵ (which may be large if the agents care much more about the outcome than their privacy) at which the privacy cost is small enough relative to the benefit that agents are willing to report truthfully. Moreover, by taking agents' value for privacy into account in the incentives analysis, we can have greater confidence that the agents will actually report truthfully and achieve the approximately optimal social welfare our analysis predicts.

The appropriate setting of ϵ depends on agents' relative valuation of privacy versus the outcomes, and needs to be determined exogenously to apply our mechanisms. Eliciting agents' values for privacy is an orthogonal problem to the one we consider, and is the subject of a different line of work, starting with Ghosh and Roth [2011] (discussed more in the next section). Since we use a dominant-strategy solution concept, it is OK if our assumptions about the privacy valuations and hence our setting of ϵ are incorrect for some of the agents; the other agents will still have an incentive to report truthfully.

1.3. Other Related Work

Independently of our work, Nissim, Orlandi, and Smorodinsky [2011] have considered a related way of modelling privacy in players' utilities and constructed truthful mechanisms under their model. They assume that if *all* outcomes o have the property that no player's report affects the probability of o much (i.e., the mechanism is differentially private), then the *overall* privacy cost of the mechanism is small for every player. This is weaker than our assumption, which requires an analogous bound on the privacy cost for each specific outcome o. Indeed, Nissim et al. [2011] do not consider a per-outcome model of privacy, and thus do not obtain a reduced privacy cost when a player has a very low probability of affecting the outcome (e.g., a highly skewed election). However, they require assumptions that give the mechanism an ability to reward players for truthfulness (through their concept of "agents' reactions," which can be restricted by the mechanism). For example, in the case of an election or poll between two choices (also considered in their paper), they require that a player *directly* benefits from reporting their true choice (e.g., in a poll to determine which of two magazines is more popular, a player will receive a copy of whichever magazine she votes for, providing her with positive utility at no privacy cost), whereas we consider a more standard election where the players only receive utility for their preferred candidate winning (minus any costs due to privacy). In general, we consider the standard mechanism design setting where truthfulness is only rewarded through the public outcome (and possibly payments), and this brings out the main tension discussed earlier: we can only incentivize truthfulness by giving players an influence on the outcome, but such an influence also leads to privacy costs, which may incentivize lying.

Another recent paper that considers a combination of differential privacy and mechanism design is that of Ghosh and Roth [2011]. They consider a setting where each player has some private information and some value for her privacy (measured in a way related to differential privacy). The goal is to design a mechanism for a data analyst to compute a statistic of the players' private information as accurately as possible, by purchasing data from many players and then performing a differentially private computation. In their model, players may lie about their value for privacy, but they cannot provide false data to the analyst. So they design mechanisms that get players to truthfully report their value for privacy. In contrast, we consider settings where players may lie about their data (their private types), but where they have a direct interest in the outcome of the mechanism, which we use to outweigh their value for privacy (so we do not need to explicitly elicit their value for privacy).

Subsequent to our work, Huang and Kannan [2012] examined the properties of the exponential mechanism [2007], which can be thought of as noisy version of VCG that is slightly different from the one we study. They showed that, with appropriate payments, this mechanism is truthful, individually rational, approximately efficient, and differentially private, but their model does not incorporate privacy costs into players' utility functions.

Our model assumes that while the outcome selected by the mechanism can be observed, the actions of the players cannot. Gradwohl [2012] explores implementation with players with privacy preferences in settings where the actions can be observed and so techniques based on differential privacy are not a natural fit. Gradwohl and Smorodinsky [2014] consider a setting where an others make inferences about a player who was preferences over their beliefs, but again these inferences are based on observed actions rather than outcomes.

For other related and subsequent work, there have been several surveys of mechanism design and differential privacy [Pai and Roth 2013; Dwork and Roth 2014].

We remark that there have also been a number of works that consider secure-computation-like notions of privacy for mechanism design problems (see [Naor et al. 1999; Dodis et al. 2000; Izmalkov et al. 2005; Parkes et al. 2008; Brandt and Sandholm 2008; Feigenbaum et al. 2010] for some examples). In these works, the goal is to ensure that a distributed implementation of a mechanism does not leak much more information than a centralized implementation by a trusted third party In our setting, we assume we have a trusted third party to implement the mechanism and are concerned with the information leaked by the outcome itself.

2. BACKGROUND ON MECHANISM DESIGN

In this section, we introduce the standard framework of mechanism design to lay the ground for modelling privacy in the context of mechanism design in next section. We use a running example of an election between two candidates. A (deterministic) mechanism is given by the following components:

- A number n of players. These might be the n voters in an election between two candidates A and B.
- A set Θ of player types. In the election example, we take $\Theta = \{A, B\}$, where $\theta_i \in \Theta$ indicates which of the two candidates is preferred by voter $i \in [n]$.
- A set O of outcomes. In the election example, we take $O = \{A, B\}$, where the outcome indicates which of the two candidates win. (Note that we do not include the tally of the vote as part of the outcome. This turns out to be significant for privacy.)
- Players' action spaces X_i for all $i \in [n]$. In general, a player's action space can be different from his type space. However, in this paper we view the types in Θ to be values that we expect players to know and report. Hence, we require $X_i = \Theta$ for all $i \in [n]$ (i.e., we restrict to direct revelation mechanisms, which is without loss of generality). In the election example, the action of a player is to vote for A or for B.
- An outcome function $\mathcal{M} : X_1 \times \cdots \times X_n \to O$ that determines an outcome given players' actions. Since we require $X_i = \Theta$, the outcome function becomes $\mathcal{M} : \Theta^n \to O$. For example, a majority voting mechanism's function maps votes of players to the candidate who received a majority of votes.
- Player-specific *utility functions* $U_i : \Theta \times O \to \mathbb{R}$ for $i = 1, \ldots, n$, giving the utility of player i as a function of his type and the outcome.

To simplify notation, we use a mechanism's outcome function to represent the mechanism. That is, a mechanism is denoted $\mathcal{M} : \Theta^n \to O$. The goal of mechanism design is then to design a *mechanism* $\mathcal{M} : \Theta^n \to O$ that takes players' (reported) types and selects an outcome so as to maximize some global objective function (e.g. the sum of the players' utilities, known as *social welfare*) even when players may falsely report their type in order to increase their personal utility. The possibility of players' misreporting is typically handled by designing mechanisms that are *incentive-compatible*, i.e., it is in each player's interest to report their type honestly. A strong formulation of incentive compatibility is the notion of *truthfulness* (a.k.a. dominant-strategy incentive compatibility): for all players i, all types $\theta_i \in \Theta$, all alternative reports $\theta'_i \in \Theta$, and all profiles θ_{-i} of the other players' reports[2], we have:

$$U_i(\theta_i, \mathcal{M}(\theta_i, \theta_{-i})) \geq U_i(\theta_i, \mathcal{M}(\theta'_i, \theta_{-i})). \tag{1}$$

If Inequality (1) holds for player i (but not necessarily all players), we say that the mechanism is *truthful for player i*. Note that we are using θ_{-i} here as both the type and the report of other players. Since truthfulness must hold for all possible reports of other players, it is without loss of generality to assume that other players report their true type. This is in contrast to the notion of a Nash equilibrium which refers to the incentives of player i under the assumption that other players are using equilibrium strategies.

[2] We adopt the standard game-theory convention that θ_{-i} refers to all components of the vector θ except the one corresponding to player i, and that (θ_i, θ_{-i}) denotes the vector obtained by putting θ_i in the i'th component and using θ_{-i} for the rest.

In the election example, it is easy to see that standard majority voting is a truthful mechanism. Changing one's vote to a less-preferred candidate can never increase one's utility (it either does not affect the outcome, or does so in a way that results in lower utility).

In this paper, we will allow randomized mechanisms, which we define as $\mathcal{M} : \Theta^n \times \mathcal{R} \to O$, where $\mathcal{R}$ is the probability space from which the mechanism makes its random choices (e.g., all possible sequences of coin tosses used by the mechanism). We write $\mathcal{M}(\theta)$ to denote the random variable obtained by sampling r from $\mathcal{R}$ and evaluating $\mathcal{M}(\theta; r)$. This (non-standard) definition of a randomized mechanism is equivalent to the standard one (where the mechanism is a function from reported types to a distribution over outcomes) and makes our analysis clearer.

For randomized mechanisms, one natural generalization of truthfulness is *truthfulness in expectation*: for all players i, all types θ_i, all utility functions U_i, all reports θ'_i, and all profiles θ_{-i} of the other players' reports, we have:

$$\mathrm{E}[U_i(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}))] \geq \mathrm{E}[U_i(\theta_i, \mathcal{M}(\theta'_i, \theta_{-i}))],$$

where the expectation is taken over the random choices of the mechanism.

A stronger notion is that of *universal truthfulness*: for all players i, all types θ_i and utility functions U_i, all alternative reports θ'_i, and all profiles θ_{-i} of the other players' reports, and all $r \in \mathcal{R}$, we have:

$$U_i(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r)) \geq U_i(\theta_i, \mathcal{M}(\theta'_i, \theta_{-i}; r)).$$

Thus $\mathcal{M}$ being universally truthful is equivalent to saying that for every $r \in \mathcal{R}$, $\mathcal{M}(\cdot; r)$ is a deterministic truthful mechanism.

3. MODELLING PRIVACY IN MECHANISM DESIGN

The standard framework of mechanism design does not consider a player's value of privacy. In this section, we incorporate privacy into mechanism design and adapt the definitions of truthfulness accordingly. We continue considering the basic mechanism-design setting from Section 2. However, players now care not only about the outcome of the mechanism, but also what that outcome reveals about their private types. Thus, a player's utility becomes

$$U_i = U_i^{out} + U_i^{priv}, \tag{2}$$

where $U_i^{out} : \Theta \times O \to \mathbb{R}$ is player i's utility for the outcome and U_i^{priv} is player i's utility associated with privacy or information leakage. Before discussing the form of U^{priv} (i.e., what are its inputs), we note that in Equation (2), there is already an implicit assumption that privacy can be measured in units that can be linearly traded with other forms of utility. A more general formulation would allow U_i to be an arbitrary monotone function of U_i^{out} and U_i^{priv}, but we make the standard quasi-linearity assumption for simplicity.

Now, we turn to functional form of U_i^{priv}. First, we note that U_i^{priv} should not just be a function of player i's type and the outcome. What matters is the *functional relationship* between player i's reported type and the outcome. For example, a voting mechanism that ignores player i's vote should have zero privacy cost to player i, but one that uses player i's vote to entirely determine the outcome may have a large privacy cost. So we will allow U_i^{priv} to depend on the mechanism itself, as well as the reports of other players, since these are what determine the functional relationship between player i's report and the outcome:

$$U_i^{priv} : \Theta \times O \times \{\mathcal{M} : \Theta^n \times \mathcal{R} \to O\} \times \Theta^{n-1} \to \mathbb{R}. \tag{3}$$

Thus, when the reports of the n players are $\theta' \in \Theta^n$ and the outcome is o, the utility of player i is

$$U_i(\theta_i, o, \mathcal{M}, \theta'_{-i}) = U_i^{out}(\theta_i, o) + U_i^{priv}(\theta_i, o, \mathcal{M}, \theta'_{-i}).$$

In particular, U_i has the same inputs as U^{priv} above, including $\mathcal{M}$. Unlike standard mechanism design, we are not given fixed utility functions and then need to design a mechanism with respect to those utility functions. Our choice of mechanism affects the utility functions too! This has implications for the revelation principle, which we discuss in Section 9.

Note that we do not assume that U_i^{priv} is always negative (in contrast to Xiao [2011]). In some cases, players may prefer for information about them to be kept secret and in other cases they may prefer for it to be leaked (e.g., in case it is flattering). Thus, U_i^{priv} may be better thought of as "informational utility" rather than a "privacy cost".

It is significant that we do not allow the U_i^{priv} to depend on the *report* or, more generally, the *strategy* of player i. This is again in contrast to Xiao's modelling of privacy [Xiao 2011]. We will discuss the motivation for our choice in Section 8, and also show that despite this difference, truthfulness with respect to our modelling implies truthfulness with respect to Xiao's modelling (Appendix A).

Clearly no mechanism design would be possible if we make no further assumptions about the U_i^{priv}'s and allow them to be arbitrary, unknown functions (as their behavior could completely cancel the U_i^{out}'s). Thus, we will make the natural assumption that U_i^{priv} is small if player i's report has little influence on the outcome o. More precisely:

ASSUMPTION 3.1 (PRIVACY-VALUE ASSUMPTION).

$$\forall \theta \in \Theta^n, o \in O, \mathcal{M} : \left| U_i^{priv}(\theta_i, o, \mathcal{M}, \theta_{-i}) \right| \le F_i \left(\max_{\theta_i', \theta_i'' \in \Theta} \frac{\Pr\left[\mathcal{M}(\theta_i', \theta_{-i}) = o\right]}{\Pr\left[\mathcal{M}(\theta_i'', \theta_{-i}) = o\right]} \right),$$

where $F_i : [1, \infty) \to [0, \infty]$ *is a* privacy-bound *function with the property that* $F_i(x) \to 0$ *as* $x \to 1$*, and the probabilities are taken over the random choices of* $\mathcal{M}$*.*

Note that if the mechanism ignores player i's report, then the right-hand side of (3.1) is $F_i(1)$, which naturally corresponds to a privacy cost of 0. Thus, we are assuming that the privacy costs satisfy a continuity condition as the mechanism's dependence on player i's report decreases. The privacy-bound function F_i could be the same for all players, but we allow it to depend on the player for generality.

Assumption (3.1) is inspired by the notion of *differential privacy*, which is due to Dinur and Nissim [2003], Dwork and Nissim [2004], Blum et al. [2005], and Dwork et al. [2006]. We restate it in our notation:

Definition 3.2. A mechanism $\mathcal{M} : \Theta^n \times \mathcal{R} \to O$ is *ϵ-differentially private* iff

$$\forall \theta_{-i} \in \Theta^{n-1}, o \in O \qquad \max_{\theta_i', \theta_i'' \in \Theta} \frac{\Pr\left[\mathcal{M}(\theta_i', \theta_{-i}) = o\right]}{\Pr\left[\mathcal{M}(\theta_i'', \theta_{-i}) = o\right]} \le e^{\epsilon}.$$

By inspection of Assumption (3.1) and the definition of differential privacy, we have the following result.

PROPOSITION 3.3. *If* $\mathcal{M}$ *is* ϵ*-differentially private, then for all players i whose utility functions satisfy Assumption (3.1), all* $\theta_{-i} \in \Theta^{n-1}$*, and* $o \in O$*, we have* $\left| U_i^{priv}(\theta_i, o, \mathcal{M}, \theta_{-i}) \right| \le F_i(e^{\epsilon})$*.*

In particular, as we take $\epsilon \to 0$, the privacy cost of any given outcome tends to 0.

Like differential privacy, Assumption (3.1) makes sense only for randomized mechanisms, and only measures the loss in privacy contributed by Player i's report when fixing the reports of the other players. In some cases, it may be that the other players' reports already reveal a lot of information about player i. See Section 8 for further discussion, interpretation, and critiques of our modelling. With this model, the definitions of truthfulness with privacy are direct analogues of the basic definitions given earlier.

Definition 3.4 (*truthfulness with privacy*). Consider a mechanism design problem with n players, type space Θ, and outcome space O. For a player i with utility function $U_i = U_i^{out} + U_i^{priv}$, we say that a randomized mechanism $\mathcal{M} : \Theta^n \times \mathcal{R} \to O$ is *truthful in expectation for player i* if for all types $\theta_i \in \Theta_i$, all alternative reports $\theta_i' \in \Theta$ for player i, and all possible profiles θ_{-i} of the other players' reports, we have:

$$\mathrm{E}[U_i(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}, \theta_{-i})] \geq \mathrm{E}[U_i(\theta_i, \mathcal{M}(\theta_i', \theta_{-i}), \mathcal{M}, \theta_{-i})].$$

We say that $\mathcal{M}$ is *universally truthful for player i* if the inequality further holds for all values of $r \in \mathcal{R}$:

$$U_i(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r), \mathcal{M}, \theta_{-i}) \geq U_i(\theta_i, \mathcal{M}(\theta_i', \theta_{-i}; r), \mathcal{M}, \theta_{-i}).$$

Note that, unlike in standard settings, $\mathcal{M}$ being universally truthful does *not* mean that the deterministic mechanisms $\mathcal{M}(\cdot; r)$ are truthful. Indeed, even when we fix r, the privacy utility $U_i^{priv}(\theta, o, \mathcal{M}, \theta_{-i})$ still depends on the original randomized function $\mathcal{M}$, and the privacy properties of $\mathcal{M}$ would be lost if we publicly revealed r. What universal truthfulness means is that player i would still want to report truthfully even if she knew r but it were kept secret from the rest of the world.

4. PRIVATE TWO-CANDIDATE ELECTIONS

Using Proposition 3.3, we will sometimes be able to obtain truthful mechanisms taking privacy into account by applying tools from differential privacy to mechanisms that are already truthful when ignoring privacy. In this section, we give an illustration of this approach in our example of a two-candidate election.

MECHANISM 4.1. *Differentially private election mechanism*
Input: profile $\theta \in \{A, B\}^n$ of votes, privacy parameter $\epsilon > 0$.

(1) *Choose $r \in \mathbb{Z}$ from a discrete Laplace distribution, namely* $\Pr[r = k] \propto \exp(-\epsilon|k|)$.
(2) *If $\#\{i : \theta_i = A\} - \#\{i : \theta_i = B\} \geq r$, output A. Otherwise output B.*

We show that for sufficiently small ϵ, this mechanism is truthful for players satisfying Assumption 3.1:

THEOREM 4.2. *Mechanism 4.1 is universally truthful for player i provided that, for some function F_i:*

(1) *Player i's privacy utility U_i^{priv} satisfies Assumption 3.1 with privacy bound function F_i, and*
(2) $U_i^{out}(\theta_i, \theta_i) - U_i^{out}(\theta_i, \neg\theta_i) \geq 2F_i(e^\epsilon)$,

Note that Condition 2 holds for sufficiently small $\epsilon > 0$ (since $F_i(x) \to 0$ as $x \to 1$). The setting of ϵ needed to achieve truthfulness depends only on how much the players value their preferred candidate (measured by the left-hand side of Condition 2) and how much they value privacy (measured by the right-hand side of Condition 2), and is independent of the number of players n.

PROOF. Fix the actual type $\theta_i \in \{A, B\}$ of player i, a profile θ_{-i} of reports of the other players, and a choice r for $\mathcal{M}$'s randomness. The only alternate report for player i we need to consider is $\theta_i' = \neg\theta_i$. Let $o = \mathcal{M}(\theta_i, \theta_{-i}; r)$ and $o' = \mathcal{M}(\neg\theta_i, \theta_{-i}; r)$. We need to show that $U_i(\theta_i, o, \mathcal{M}, \theta_{-i}) \geq U_i(\theta_i, o', \mathcal{M}, \theta_{-i})$, or equivalently

$$U_i^{out}(\theta_i, o) - U_i^{out}(\theta_i, o') \geq U_i^{priv}(\theta_i, o', \mathcal{M}, \theta_{-i}) - U_i^{priv}(\theta_i, o, \mathcal{M}, \theta_{-i}). \tag{4}$$

We consider two cases:

Case 1: $o = o'$. In this case, Inequality (4) holds because both the left-hand and right-hand sides are zero.

Case 2: $o \neq o'$. This implies that $o = \theta_i$ and $o' = \neg\theta_i$. (If player i's report has any effect on the outcome of the differentially private voting mechanism, then it must be that the outcome equals player i's report.) Thus the left-hand side of Inequality (4) equals $U_i^{out}(\theta_i, \theta_i) - U_i^{out}(\theta_i, \neg\theta_i)$. By

Proposition 3.3, the right-hand side of Inequality (4) is at most $2F(e^{\epsilon})$. Thus, Inequality (4) holds by hypothesis.

□

Of course, truthfulness is not the only property of interest. After all, a mechanism that is simply a constant function is (weakly) truthful. Another property we would like is economic *efficiency*. Typically, this is defined as maximizing social welfare, the sum of players' utilities. Here we consider the sum of *outcome* utilities for simplicity. As is standard, we normalize players' utilities so that all players are counted equally in measuring the social welfare. In our voting example, we wish to maximize the number of voters' whose preferred candidates win, which is equivalent to normalizing the left-hand side of Condition 2 in Theorem 4.2 to 1. Standard, deterministic majority voting clearly maximizes this measure of social welfare. Our mechanism achieves approximate efficiency:

Proposition 4.3. *For every profile $\theta \in \Theta^n$ of reports, if we select $o \leftarrow \mathcal{M}(\theta)$ using Mechanism 4.1, then:*

(*1*) $\Pr\left[\#\{i : \theta_i = o\} \leq \max_{o' \in \{A,B\}} \#\{i : \theta_i = o'\} - \Delta\right] < e^{-\epsilon\Delta}$.
(*2*) $\mathrm{E}\left[\#\{i : \theta_i = o\}\right] > \max_{o' \in \{A,B\}} \#\{i : \theta_i = o'\} - 1/\epsilon$.

Proof. The maximum number of voters will be satisfied by taking the majority candidate $o^* = \mathrm{Maj}(\theta)$, where we break ties in favor of A. Let $\Delta' = \#\{i : \theta_i = o^*\} - \#\{i : \theta_i = \neg o^*\}$. If $o^* = A$, then $\neg o^* = B$ is selected iff the noise r is larger than Δ'. If $o^* = B$, then $\neg o^* = A$ is selected iff the noise r is smaller than or equal to $-\Delta'$. Since r is chosen so that $\Pr[r = k] \propto e^{-\epsilon|k|}$, the probability of selecting $\neg o^*$ in either case is bounded as:

$$\Pr[\mathcal{M}(\theta) = \neg o^*] \leq \frac{\sum_{k \geq \Delta'} e^{-\epsilon k}}{\sum_{k \in \mathbb{Z}} e^{-\epsilon|k|}} = \frac{e^{-\epsilon\Delta'}}{1 + e^{-\epsilon}} \leq e^{-\epsilon\Delta'}.$$

Now the high probability bound follows by considering the case that $\Delta' \geq \Delta$ (otherwise the event occurs with probability 0). The expectation bound is computed as follows:

$$\begin{aligned} &\mathrm{E}\left[\max_{o' \in \{A,B\}} \#\{i : \theta_i = o'\} - \#\{i : \theta_i = \mathcal{M}(\theta)\}\right] \\ &= \Pr[\mathcal{M}(\theta) = \neg o^*] \cdot \Delta' \leq \Delta' \cdot \tfrac{e^{-\epsilon\Delta'}}{1+e^{-\epsilon}} \leq \tfrac{1}{e\epsilon} \cdot \tfrac{1}{1+e^{-\epsilon}} < \tfrac{1}{\epsilon}, \end{aligned}$$

where the second-to-last inequality follows from the fact that $xe^{-\epsilon x}$ is minimized at $x = 1/\epsilon$. □

Thus, the number of voters whose preferred candidate wins is within $O(1/\epsilon)$ of optimal, in expectation and with high probability. This deviation is independent of n, the number of players. Thus if we take ϵ to be a constant (as suffices for truthfulness) and let $n \to \infty$, the economic efficiency approaches optimal, when we consider both as fractions of n. This also holds for vanishing $\epsilon = \epsilon(n)$, provided $\epsilon = \omega(1/n)$. Despite the notation, ϵ need not be tiny, which allows a tradeoff between efficiency and privacy.

This analysis considers the social welfare as a sum of outcome utilities (again normalizing so that everyone values their preferred candidate by one unit of utility more than the other candidate). We can consider the effect of privacy utilities on the social welfare too. By Proposition 3.3, the privacy utilities affect the social welfare by at most $\sum_i F_i(e^{\epsilon})$, assuming player i satisfies Assumption 3.1 with privacy bound function F_i. If all players satisfy Assumption 3.1 with the same privacy bound function $F_i = F$, then the effect on social welfare is at most $n \cdot F(e^{\epsilon})$. By taking $\epsilon \to 0$ (e.g., $\epsilon = 1/\sqrt{n}$), the privacy utilities contribute a vanishing fraction of n.

Another desirable property is *individual rationality*: players given the additional option of not participating should still prefer to participate and report truthfully. This property follows from the same argument we used to establish universal truthfulness. By dropping out, the only change in outcome that player i can create is to make her less preferred candidate win. Thus, the same argument as in Theorem 4.2 shows that player i prefers truthful participation to dropping out.

Proposition 4.4. *Under the same assumptions as Theorem 4.2, Mechanism 4.1 is individually rational for player i.*

5. TOOLS FOR PROVING TRUTHFULNESS WITH PRIVACY

The analysis of truthfulness in Theorem 4.2 is quite general. It holds for any differentially private mechanism with the property that if a player can actually change the outcome of the mechanism by reporting untruthfully, then it will have a noticeable negative impact on the player's outcome utility. We abstract this property for use in analyzing our other mechanisms.

Lemma 5.1. *Consider a mechanism design problem with n players, type space Θ, and outcome space O. Let player i have a utility function $U_i = U_i^{out} + U_i^{priv}$ satisfying Assumption 3.1 with privacy bound function F_i. Suppose that randomized mechanism $\mathcal{M} : \Theta^n \to O$ has the following properties:*

(1) $\mathcal{M}$ is ϵ-differentially private, and
(2) For all possible types θ_i, all profiles θ_{-i} of the other players' reports, all random choices r of $\mathcal{M}$, and all alternative reports θ_i' for player i: if $\mathcal{M}(\theta_i, \theta_{-i}; r) \neq \mathcal{M}(\theta_i', \theta_{-i}; r)$, then $U^{out}(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r)) - U^{out}(\theta_i, \mathcal{M}(\theta_i', \theta_{-i}; r)) \geq 2F_i(e^\epsilon)$,

Then $\mathcal{M}$ is universally truthful for player i.

Lemma 5.1 implicitly requires that players not be indifferent between outcomes (unless there is no way for the player to change one outcome to the other). A condition like this is necessary because otherwise players may be able to find reports that have no effect on their outcome utility but improve their privacy utility. However, we show in Section 7 that in some settings with payments, truthfulness can be achieved even with indifference between outcomes since payments can break ties.

It is also illustrative and useful to consider what happens when we take the expectation over the mechanism's coin tosses. We can upper-bound the privacy utility as follows:

Lemma 5.2. *Consider a mechanism design problem with n players, type space Θ, and outcome space O. Let player i have type $\theta_i \in \Theta_i$ and a utility function $U_i = U_i^{out} + U_i^{priv}$ satisfying Assumption 3.1. Suppose that randomized mechanism $\mathcal{M} : \Theta^n \times \mathcal{R} \to O$ is ϵ-differentially private. Then for all possible profiles θ_{-i} of the other players' reports, all random choices r of $\mathcal{M}$, and all alternative reports θ_i' for player i, we have*

$$\left|\mathrm{E}[U_i^{priv}(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}, \theta_{-i})] - \mathrm{E}[U_i^{priv}(\theta_i, \mathcal{M}(\theta_i', \theta_{-i}), \mathcal{M}, \theta_{-i})]\right| \leq 2F_i(e^\epsilon)\cdot SD(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta_i', \theta_{-i})),$$

where SD denotes statistical difference.[3]

Proof. For every two discrete random variables X and Y taking values in a universe $\mathcal{U}$, and every function $f : \mathcal{U} \to [-1, 1]$, it holds that $|\mathrm{E}[f(X)] - \mathrm{E}[f(Y)]| \leq 2\mathrm{SD}(X, Y)$. (The f that maximizes the left-hand side sets $f(x) = 1$ when $\Pr[X = x] > \Pr[Y = x]$ and sets $f(x) = -1$ otherwise.) Take $\mathcal{U} = O$, $X = \mathcal{M}(\theta_i, \theta_{-i})$, $Y = \mathcal{M}(\theta_i', \theta_{-i})$, and $f(o) = U_i^{priv}(\theta_i, o, \mathcal{M}, \theta_{-i})/F_i(e^\epsilon)$. By Proposition 3.3, we have $f(o) \in [-1, 1]$. □

By this lemma, to establish truthfulness in expectation, it suffices to show that the expected gain in outcome utility from reporting θ_i instead of θ_i' grows proportionally with $\mathrm{SD}(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta_i', \theta_{-i}))$. (Specifically, it should be at least the statistical difference times $2F_i(e^\epsilon)$.) In Lemma 5.1, the gain in outcome utility is related to the statistical difference by coupling the random variables $\mathcal{M}(\theta_i, \theta_{-i})$ and $\mathcal{M}(\theta_i', \theta_{-i})$ according to the random choices r of $\mathcal{M}$. Indeed, $\Pr_r[\mathcal{M}(\theta_i, \theta_{-i}; r) \neq \mathcal{M}(\theta_i', \theta_{-i}; r)] \geq \mathrm{SD}(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta_i', \theta_{-i}))$. Thus, if the outcome-utility gain from truthfulness is larger than $2F_i(e^\epsilon)$ whenever $\mathcal{M}(\theta_i, \theta_{-i}; r) \neq \mathcal{M}(\theta_i', \theta_{-i}; r)$, then we have truthfulness in expectation (indeed, even universal truthfulness).

[3] The *statistical difference* (aka *total variation distance*) between two discrete random variables X and Y taking values in a universe $\mathcal{U}$ is defined to be $\mathrm{SD}(X, Y) = \max_{S \subseteq \mathcal{U}} |\Pr[X \in S] - \Pr[Y \in S]|$.

We note that if $\mathcal{M}$ is differentially private, then $\mathrm{SD}(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta'_i, \theta_{-i})) \le e^\epsilon - 1 = O(\epsilon)$, for small ϵ. By Lemma 5.2, the expected difference in privacy utility between any two reports is at most $O(F_i(e^\epsilon) \cdot \epsilon)$. Thus, ϵ-differential privacy helps us twice, once in bounding the pointwise privacy cost (as $2F_i(e^\epsilon)$), and second in bounding the statistical difference between outcomes. On the other hand, for mechanisms satisfying the conditions of Lemma 5.1, the differential privacy only affects the expected outcome utility by a factor related to the statistical difference. This is why, by taking ϵ sufficiently small, we can ensure that the outcome utility of truthfulness dominates the privacy cost.

Lemma 5.2 is related to, indeed inspired by, existing lemmas used to analyze the composition of differentially private mechanisms. These lemmas state that while differential privacy guarantees a worst case bound of ϵ on the "privacy loss" of all possible outputs, this actually implies an *expected* privacy loss of $O(\epsilon^2)$. Such bounds correspond to the special case of Lemma 5.2 when $F_i = \ln$ and we replace the statistical difference with the upper bound $e^\epsilon - 1$. These $O(\epsilon^2)$ bounds on expected privacy loss were proven first in the case of specific mechanisms by Dinur and Nissim [2003] and Dwork and Nissim [2004], and then in the case of arbitrary differentially private mechanisms by Dwork et al. [2010]. In our case, the $O(\epsilon^2)$ bound does not suffice, and we need the stronger bound expressed in terms of the statistical difference. Consider the differentially private election when the vote is highly skewed (e.g., 2/3 vs. 1/3). Then a player has only an exponentially small probability (over the random choice r of the mechanism) of affecting the outcome, and so the expected outcome utility for voting truthfully is exponentially small. On the other hand, by Lemma 5.2, the expected privacy loss is also exponentially small, so we can still have truthfulness.

6. DISCRETE FACILITY LOCATION

In this section, we apply our framework to discrete facility location. Let $\Theta = \{\ell_1 < \ell_2 < \cdots < \ell_q\} \subset [0, 1]$ be a finite set of types indicating player's preferred locations for a facility on the unit interval and $O = [0, 1]$. Players prefer to have the facility located as close to them as possible: $U_i^{out}(\theta_i, o) = -|\theta_i - o|$. For example, the mechanism may be selecting a location for a bus stop along a major highway, and the locations $\ell_1, \ldots, \ell_q$ might correspond to cities along the highway where potential bus riders live.

Note that the voting game we previously considered can be represented as the special case where $\Theta = \{0, 1\}$. This problem has a well-known truthful and economically efficient mechanism: select the location of the median report. Xiao [2011] gave a private and truthful mechanism for this problem based on taking the median of a perturbed histogram. His analysis only proved that the mechanism satisfies "approximate differential privacy" (often called (ϵ, δ) differential privacy). To use Proposition 3.3, we need the mechanism to satisfy pure ϵ differential privacy (as in Definition 3.2). Here we do that for a variant of Xiao's mechanism.

MECHANISM 6.1. *Differentially private discrete facility location mechanism*
Input: profile $\theta \in \Theta^n$ of types, privacy parameter $\epsilon > 0$.

(1) *Construct the histogram $h = (h_1, \ldots, h_q)$ of reported type frequencies where h_j is the number of reports θ_i of type ℓ_j and $q = |\Theta|$.*
(2) *Choose a random (nonnegative, integer) noise vector $r = (r_1, \ldots, r_q) \in \mathbb{N}^q$ where the components r_j are chosen independently such that $\Pr\left[r_j = k\right]$ is proportional to $\exp(-\epsilon k/2)$.*
(3) *Output the type corresponding to median of the perturbed histogram $h + r$. That is, we output $\ell_{\mathrm{Med}(h+r)}$, where for $z \in \mathbb{N}^q$ we define $\mathrm{Med}(z)$ to be the minimum $k \in [q]$ such that $\sum_{j=1}^k z_j \ge \sum_{j=k+1}^q z_j$).*

Xiao's mechanism instead chooses the noise components r_j according to a truncated and shifted Laplace distribution. Specifically, $\Pr[r_j = k]$ is proportional to $\exp((\epsilon/2) \cdot |k - t|)$ for $k = 0, \ldots, 2t$ and $\Pr[r_j = k] = 0$ for $k > 2t$, where $t = \Theta(\log(1/\delta)/\epsilon)$. This ensures that the noisy histogram $h + r$ is $(\epsilon, q\delta)$ differentially private, and hence the outcome $\ell_{\mathrm{Med}(h+r)}$ is as well. Our proof directly analyzes the median, without passing through the histogram. This enables us to achieve pure ϵ differential privacy and use a simpler noise distribution. On the other hand, Xiao's analysis is more general, in that it applies to any mechanism that computes its result based on a noisy histogram.

Lemma 6.2. *Mechanism 6.1 is ϵ-differentially private.*

Proof. Differential privacy requires that on any pair of histograms h, h' reachable by one player reporting different types, the probability of any particular outcome $o = \ell_j$ being selected differs by at most an e^ϵ multiplicative factor. Since reporting a different type results in two changes to the histogram (adding to one type and subtracting from another), we show that on each such change the probability differs by at most an $e^{\epsilon/2}$ factor.

Consider two histograms h and h' that differ only by an addition or subtraction of 1 to a single entry. Let $f_j : \mathbb{N}^q \to \mathbb{N}^q$ map a vector s to the vector $(s_j + 1, s_{-j})$ (i.e., identical except s_j has been increased by 1). f_j is an injection and has the property that if j is the median of $h+s$ then j is also the median of $h' + f_j(s)$. Note that under our noise distribution, we have $\Pr[r = s] = e^{\epsilon/2} \cdot \Pr[r = f_j(s)]$.

Then writing $\mathcal{M}$ as a function of h rather than θ, we have:

$$\begin{aligned}\Pr\left[\mathcal{M}(h) = \ell_j\right] &= \sum_{s \text{ s.t. } \mathrm{Med}(h+s)=j} \Pr\left[r = s\right] \\ &= \sum_{s \text{ s.t. } \mathrm{Med}(h+s)=j} e^{\epsilon/2} \cdot \Pr\left[r = f_j(s)\right] \\ &\leq e^{\epsilon/2} \cdot \sum_{s \text{ s.t. } \mathrm{Med}(h'+f_j(s))=j} \Pr\left[r = f_j(s)\right] \\ &\leq e^{\epsilon/2} \sum_{s' \text{ s.t. } \mathrm{Med}(h'+s')=j} \Pr\left[r = s'\right] \\ &= e^{\epsilon/2} \cdot \Pr\left[\mathcal{M}(h') = \ell_j\right].\end{aligned}$$

A symmetric argument shows this is also true switching h and h', which completes the proof. □

We note that the only property the proof uses about the noise distribution is that $\Pr[r = s] = e^{\epsilon/2} \cdot \Pr[r = f_j(s)]$. This property does not hold for Xiao's noise distribution as described, due to it being truncated above at $2t$, but would hold if his noise distribution was truncated only below.

We next show that this mechanism is truthful and individually rational.

Theorem 6.3. *Mechanism 6.1 is universally truthful and individually rational for player i provided that, for some function F_i:*

(1) *Player i's privacy utility U_i^{priv} satisfies Assumption 3.1 with privacy bound function F_i, and*
(2) *For all $o, o' \in \Theta$ such that $\theta_i < o < o'$ or $o' > o > \theta_i$, we have $U_i^{out}(\theta_i, o) - U_i^{out}(\theta_i, o') \geq 2F_i(e^\epsilon)$.*

In particular, if all players share the standard outcome utility function $U_i^{out}(\theta_i, o) = -|\theta_i - o|$ and have the same privacy bound function $F_i = F$, then the mechanism is universally truthful and individually rational provided that

$$\min_{j \neq k} |\ell_j - \ell_k| \geq 2F(e^\epsilon).$$

So, for a fixed set Θ of player types (preferred locations), we can take ϵ to be a small constant and have truthfulness and individual rationality.

Proof.

Fix $r \in \mathbb{N}^q$, the randomness used by the mechanism and the reports θ_{-i} of other players. Following Xiao [2011], we think of r as representing the reports of some fictional additional players, and follow the truthfulness reasoning for the standard, noiseless median mechanism. Suppose $\mathcal{M}(\theta_i, \theta_{-i}; r) = o$ and $\mathcal{M}(\theta_i', \theta_{-i}; r) = o' \neq o$. If $\theta_i < o$, then no other report of player i can reduce the median, so we must have $o' > o$. Thus, this change has moved the facility at least one location away from i's preferred location. Similarly, if $\theta_i > o$, we have $o' < o$ so again the change is away from i's preferred location. Therefore, universal truthfulness follows by Lemma 5.1. For individual rationality, we can model non-participation as a report of a type $\perp$ that does not get included in the histogram. Again, any change of of the median caused by reporting $\perp$ will move it away from i's preferred location. Thus $\mathcal{M}$ is individually rational. □

Proposition 6.4. *Suppose that every player i has the standard outcome utility function $U_i^{out}(\theta_i, o) = -|\theta_i - o|$. Then for every profile of types $\theta \in \Theta^n$, if we choose $o \leftarrow \mathcal{M}(\theta)$ using Mechanism 6.1, we have*

(*1*) $\Pr\left[\sum_i U_i^{out}(\theta_i, o) \le \max_{o'}\left(\sum_i U_i^{out}(\theta_i, o')\right) - \Delta\right] \le q \cdot e^{-\epsilon\Delta/q}$.

(*2*) $\mathrm{E}\left[\sum_i U_i^{out}(\theta_i, o)\right] \ge \max_{o'}\left(\sum_i U_i^{out}(\theta_i, o')\right) - O(q/\epsilon)$.

Thus, the social welfare is within $\Delta = \tilde{O}(q)/\epsilon$ of optimal, both in expectation and with high probability. Like with Proposition 4.3, these bounds are independent of the number n of participants, so we obtain asymptotically optimal social welfare as $n \to \infty$. Also like the discussion after Proposition 4.3, by taking $\epsilon = \epsilon(n)$ to be such that $\epsilon = o(1)$ and $\epsilon = \omega(1/n)$ (e.g., $\epsilon = 1/\sqrt{n}$), the sum of privacy utilities is a vanishing fraction of n (for participants satisfying Assumption 3.1 with a common privacy bound function F).

Proof. Note that $-\sum_i U_i^{out}(\theta_i, o') = \sum_j h_j \cdot |\ell_j - o'|$, where $h = (h_1, \ldots, h_q)$ is the histogram corresponding to θ. This social welfare is minimized by taking $o' = \mathrm{Med}(h)$. Our mechanism, however, computes the optimal location for the noisy histogram $h + r$. We can relate the two as follows:

$$\begin{aligned}
-\sum_i U_i^{out}(\theta_i, o) &= \sum_j h_j \cdot |\ell_j - o| \\
&\le \sum_j (h_j + r_j) \cdot |\ell_j - o| \\
&= \min_{o'} \sum_j (h_j + r_j) \cdot |\ell_j - o'| \\
&\le \min_{o'} \sum_j h_j \cdot |\ell_j - o'| + \sum_j r_j \\
&= -\max_{o'} \sum_i U_i^{out}(\theta_i, o') + \sum_j r_j.
\end{aligned}$$

Thus, for the high probability bound, it suffices to bound the probability that $\sum_j r_j \ge \Delta$. This in turn is bounded by q times the probability that any particular r_j is at least Δ/q, which is at most $e^{-\epsilon\Delta/q}$. For the expectation bound, we have

$$\mathrm{E}[\sum_j r_j] = \sum_j \mathrm{E}[r_j] = q \cdot \frac{1}{1 - e^{-\epsilon/2}} = O\left(\frac{q}{\epsilon}\right).$$

□

7. GENERAL SOCIAL CHOICE PROBLEMS WITH PAYMENTS

In the preceding two sections we have considered social choice problems where a group needs to choose among a (typically small) set of options with mechanisms that do not use money. In this section, we apply our framework social choice problems where payments are possible using an adaptation of the Vickrey-Clarke-Groves (VCG) mechanism. (This is the setting for which the Groves mechanism was originally designed and unlike in auction settings the number of outcomes is independent of the number of players.) In the general case we examine now, we don't assume any structure on the utility functions (other than discreteness), and thus need to use payments to incentivize players to truthfully reveal their preferences.

Specifically, the type $\theta_i \in \Theta$ of a player will specify a utility $U^{out}(\theta_i, o) \in \{0, 1, \ldots, M\}$ for each outcome o from a finite set O. This could correspond, for example, to players having values for outcomes expressible in whole dollars with some upper and lower bounds. This assumption ensures a finite set of types Θ and that if a player changes his reported value it must change by some minimum amount (1 with our particular assumption). Note that this formulation still allows players to be indifferent among outcomes. Gicen our notion of a type, all players share the same outcome utility function $U_i^{out} = U^{out}$ In order to reason about individual rationality, we also assume that the set of types includes a type $\perp$ that corresponds to not participating (i.e., $U_i^{out}(\perp, o) = 0$ for all o and i). For notational convenience, we assume that $O = \{0, 1, \ldots, |O| - 1\}$.

Our goal is to choose the outcome o^* that maximizes social welfare (ignoring privacy), i.e., $o^* = \text{argmax}_{o \in O} \sum_i U^{out}(\theta_i, o)$. A standard way to do so is the Groves mechanism, a special case of the more general VCG mechanism. Each player reports his type and then the optimal outcome o^* is chosen based on the reported types. To ensure truthfulness, each player is charged the externality he imposes on others. If $o_{-i} = \text{argmax}_o \sum_{j \neq i} U^{out}(\theta_j, o)$ is the outcome that would have been chosen without i's input, then player i makes a payment of

$$P_i = \sum_{j \neq i} \left(U^{out}(\theta_j, o_{-i}) - U^{out}(\theta_j, o^*) \right), \tag{5}$$

for a combined utility of $U^{out}(\theta_i, o^*) - P_i$.

In addition to subtracting payments from player i's utility as above, we also need to consider the effect of payments on privacy. (The modelling in Section 3 did not consider payments.) While it may be reasonable to treat the payments players make as secret, so that making the payment does not reveal information to others, the amount a player is asked to pay reveals information about the reports of *other* players. Therefore, we will require that the mechanism releases some *public* payment information π that enables all players to compute their payments, i.e., the payment P_i of player i should be a function of θ_i, π, and o^*. For example, π could just be the n-tuple $(P_1, \ldots, P_n)$, which corresponds to making all payments public. But in the VCG mechanism it suffices for π to include the value $V_o = \sum_i U^{out}(\theta_i, o)$ for all outcomes $o \in O$, since

$$P_i = (V_{o_{-i}} - U^{out}(\theta_i, o_{-i})) - (V_{o^*} - U^{out}(\theta_i, o^*)) = max_o \left((U^{out}(\theta_i, o^*) - U^{out}(\theta_i, o)) - (V_{o^*} - V_o) \right),$$

which can be computed using just the V_o's, o^*, and θ_i. Moreover, we actually only need to release the differences $V_{o^*} - V_o$, and only need to do so for outcomes o such that $V_{o^*} - V_o \leq M$, since only such outcomes have a chance of achieving the above maximum. (Recall that $U^{out}(\theta_i, o) \in \{0, 1, \ldots, M\}$.) This observation forms the basis of our mechanism, which we will show to be truthful for players that value privacy (under Assumption 3.1).

Before stating our mechanism, we summarize how we take payments into account in our modelling. Given reports $\theta' \in \Theta^n$ and randomness r, our mechanism $\mathcal{M}(\theta'; r)$ outputs a pair (o^*, π), where $o^* \in O$ is the selected outcome and π is "payment information". Each player then should send payment $P_i = P(\theta'_i, o^*, \pi)$ to the mechanism. (The payment function P is something we design together with the mechanism $\mathcal{M}$.) If player i's true type is θ_i, then her total utility is:

$$U_i(\theta_i, o^*, \pi, \mathcal{M}, \theta') = U^{out}(\theta_i, o^*) - P(\theta'_i, o^*, \pi) + U_i^{priv}(\theta_i, (o^*, \pi), \mathcal{M}, \theta'_{-i}).$$

Note that we measure the privacy of the *pair* (o^*, π), since both are released publicly.

To achieve truthfulness for players that value privacy, we will modify the VCG mechanism described above by adding noise to the values V_o. This yields the following mechanism:

MECHANISM 7.1. *Differentially private VCG mechanism*
Input: profile $\theta \in \Theta^n$ *of types, privacy parameter* $\epsilon > 0$.

(*1*) *Choose λ_o from a (discrete) Laplace distribution for each outcome o. Specifically, we set* $\Pr[\lambda_o = k] \propto \exp(-(\epsilon \cdot |k|)/(M \cdot |O|))$ *for every integer $k \in \mathbb{Z}$.*
(*2*) *Calculate values $V_o = \sum_j U^{out}(\theta_j, o) + \lambda_o + o/|O|$ for each outcome o. (Recall that we set $O = \{0, \ldots, |O| - 1\}$. The $o/|O|$ term is introduced in order to break ties.)*
(*3*) *Select outcome $o^* = \arg\max_o V_o$.*
(*4*) *Set the payment information $\pi = \{(o, V_{o^*} - V_o) : V_o \geq V_{o^*} - M\}$.*
(*5*) *Output (o^*, π).*

Each player i then sends a payment of $P_i = P(\theta_i, o^*, \pi) = \max_o \left((U^{out}(\theta_i, o^*) - U^{out}(\theta_i, o)) - (V_{o^*} - V_o)\right)$.

By standard results on differential privacy, the tuple of noisy values $\{V_o\}$ is ϵ-differentially private. Since the output (o^*, π) is a function of the V_o's, the output is also differentially private:

Lemma 7.2. *Mechanism 7.1 is ϵ-differentially private.*

We now prove that the mechanism is truthful in expectation for players that value privacy (satisfying Assumption 3.1). To do this, we use Lemma 5.2, which shows that by taking ϵ sufficiently small, the expected change in privacy utility from misreporting θ'_i instead of θ_i can be made an arbitrarily small fraction of the statistical difference $\mathrm{SD}(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta'_i, \theta_{-i}))$. Thus, to show truthfulness in expectation, it suffices to show that the statistical difference is at most a constant factor larger than the expected decrease in utility from misreporting. That is, we want to show:

$$\begin{aligned}&\mathrm{SD}(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta'_i, \theta_{-i}))\\ &\quad = O\Big(\mathrm{E}[U^{out}(\theta_i, \mathcal{M}(\theta_i, \theta_{-i})) - P(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}))] - \mathrm{E}[U^{out}(\theta_i, \mathcal{M}(\theta'_i, \theta_{-i})) - P(\theta'_i, \mathcal{M}(\theta'_i, \theta_{-i}))]\Big).\end{aligned}$$

To bound the statistical difference, we write $\mathcal{M}(\theta; r) = (\mathcal{M}^1(\theta; r), \mathcal{M}^2(\theta; r))$, where $\mathcal{M}^1$ gives the outcome o^* and $\mathcal{M}^2$ gives the payment information π. Then we have:

$$\begin{aligned}\mathrm{SD}(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta'_i, \theta_{-i})) &\leq \Pr_r[\mathcal{M}(\theta_i, \theta_{-i}; r) \neq \mathcal{M}(\theta'_i, \theta_{-i}; r)] \leq \Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r)]\\ &\quad + \Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) = \mathcal{M}^1(\theta'_i, \theta_{-i}; r) \wedge \mathcal{M}^2(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^2(\theta'_i, \theta_{-i}; r)].\end{aligned}$$

The next lemma bounds the statistical difference coming from the outcome:

Lemma 7.3.

$$\begin{aligned}\Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r)] \;&\leq |O| \cdot (\mathrm{E}[U^{out}(\theta_i, \mathcal{M}^1(\theta_i, \theta_{-i})) - P(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}))]\\ &\quad - \mathrm{E}[U^{out}(\theta_i, \mathcal{M}^1(\theta'_i, \theta_{-i})) - P(\theta'_i, \mathcal{M}(\theta_i, \theta_{-i}))]).\end{aligned}$$

Proof. It suffices to show that for every value of r, we have:

$$\begin{aligned}I[\mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r)] &\leq |O| \cdot (U^{out}(\theta_i, \mathcal{M}^1(\theta_i, \theta_{-i}; r)) - P(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r))\\ &\quad - U^{out}(\theta_i, \mathcal{M}^1(\theta'_i, \theta_{-i}; r)) - P(\theta'_i, \mathcal{M}(\theta_i, \theta_{-i}; r))),\end{aligned} \tag{6}$$

where $I[X]$ denotes the indicator for the event X. (Then taking expectation over r yields the desired result.)

If $\mathcal{M}^1(\theta_i, \theta_{-i}; r) = \mathcal{M}^1(\theta'_i, \theta_{-i}; r)$, then both the left-hand and right-hand sides are zero. (Recall that the payment made by player i on an outcome o depends only on the reports of the other players and the randomness of the mechanism.)

So consider a value of r such that $\mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r)$ (i.e., where the indicator is 1). We can treat the $\lambda_o + o/|O|$ term added to each V_o as the report of another player to the standard VCG mechanism. We know that

$$U^{out}(\theta_i, \mathcal{M}^1(\theta_i, \theta_{-i}; r)) - P(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r)) - U^{out}(\theta_i, \mathcal{M}(\theta'_i, \theta_{-i}; r)) - P(\theta'_i, \mathcal{M}(\theta_i, \theta_{-i}; r)) \geq 0$$

because VCG is incentive compatible for players who don't have a privacy utility. Since the mechanism adds an $o/|O|$ term to V_o to avoid ties, the above inequality is strict. Moreover, the left-hand side is at least $1/|O|$, which establishes Inequality (6).

In more detail, let $o^* = \mathcal{M}^1(\theta_i, \theta_{-i}; r)$ and $o' = \mathcal{M}^1(\theta'_i, \theta_{-i}; r)$ for some $o' \neq o^*$. Write $W_o = \sum_{j \neq i} U^{out}_j(\theta_j, o) + \lambda_o + o/|O|$ for each outcome o (W_o is just V_o excluding the report of player i), and $o_{-i} = \operatorname{argmax}_o W_o$. Since the mechanism chose o^* on report θ_i, we must have

$$W_{o^*} + U^{out}(\theta_i, o^*) \geq W_{o'} + U^{out}(\theta_i, o').$$

Since the fractional parts of the two sides are different multiples of $1/|O|$ (namely $o^*/|O|$ ad $o'/|O|$), we have:

$$W_{o^*} + U^{out}(\theta_i, o^*) \geq W_{o'} + U^{out}(\theta_i, o') + 1/|O|.$$

Thus:

$$\begin{aligned}
&U^{out}(\theta_i, \mathcal{M}^1(\theta_i, \theta_{-i}; r)) - P(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r))\\
&= U^{out}(\theta_i, o^*) - (W_{o_{-i}} - W_{o^*})\\
&\geq U^{out}(\theta_i, o') - (W_{o_{-i}} - W_{o'}) + 1/|O|\\
&= U^{out}(\theta_i, \mathcal{M}(\theta'_i, \theta_{-i}; r)) - P(\theta'_i, \mathcal{M}(\theta_i, \theta_{-i}; r)) + 1/|O|,
\end{aligned}$$

establishing Inequality (6). □

Now we need to prove a similar bound for the probability of misreporting only affecting the payment information π. We note that one trivial solution for handling payments is to only collect payments with a very small probability p, but increase the magnitude of the payments by a factor of $1/p$. In order for payments to not contribute more to the statistical difference than the outcome, we can take p to be the minimum possible nonzero value of the probability that a misreport can change the outcome (i.e., $\Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r)]$). However, this quantity is exponentially small in n. This would make the magnitude of payments exponentially large, which is undesirable. (Our assumption that players are risk neutral seems unreasonable in such a setting.) However, it turns out that we do not actually need to do this; our mechanism already releases payment information with sufficiently low probability. Indeed, we only release payment information relating to an outcome o when V_o is within M of V_{o^*}, and the probability that this occurs cannot be much larger than the probability that the outcome is changed from o^* to o.

LEMMA 7.4.

$$\begin{aligned}
&\Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) = \mathcal{M}^1(\theta'_i, \theta_{-i}; r) \wedge \mathcal{M}^2(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^2(\theta'_i, \theta_{-i}; r)]\\
&\leq\ 2Me^{\epsilon/|O|} \cdot \Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r)].
\end{aligned}$$

PROOF. First observe that

$$\begin{aligned}
&\Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) = \mathcal{M}^1(\theta'_i, \theta_{-i}; r) \wedge \mathcal{M}^2(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^2(\theta'_i, \theta_{-i}; r)]\\
&\leq \sum_{o_1 \neq o_2} \Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) = \mathcal{M}^1(\theta'_i, \theta_{-i}; r) = o_1 \wedge \mathcal{M}^2(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^2(\theta'_i, \theta_{-i}; r) \text{ on } o_2],
\end{aligned}$$

by which we mean that either $(o_2, V_{o_1} - V_{o_2})$ is released in one case but not the other or it is released in both cases but with different values.

Fix o_1 and o_2 as above. If $U^{out}(\theta_i, o_1) - U^{out}(\theta_i, o_2) = U^{out}(\theta'_i, o_1) - U^{out}(\theta'_i, o_2)$, then $\Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) = \mathcal{M}^1(\theta'_i, \theta_{-i}; r) = o_1 \wedge \mathcal{M}^2(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^2(\theta'_i, \theta_{-i}; r) \text{ on } o_2] = 0$ because the difference between V_{o_1} and V_{o_2} is not changed by the misreporting. So assume that $U^{out}(\theta_i, o_1) - U^{out}(\theta_i, o_2) \neq U^{out}(\theta'_i, o_1) - U^{out}(\theta'_i, o_2)$; these values must differ by at least 1 due to the discreteness assumption. Fix $\lambda_o = k_o$ for $o \neq o_2$. Denote them as a vector $\lambda_{-o_2} = k_{-o_2}$. Consider some value k_{o_2} such that when $\lambda_{o_2} = k_{o_2}$ we have $\mathcal{M}^1(\theta_i, \theta_{-i}; (k_{o_2}, k_{-o_2})) = \mathcal{M}^1(\theta'_i, \theta_{-i}; (k_{o_2}, k_{-o_2})) = o_1$ and $\mathcal{M}^2(\theta_i, \theta_{-i}; (k_{o_2}, k_{-o_2})) \neq \mathcal{M}^2(\theta'_i, \theta_{-i}; (k_{o_2}, k_{-o_2}))$ on o_2. (If there is no such k_{o_2} then the event has probability 0 for this choice of k_{-o_2}.) Now consider increasing the value of λ_{o_2}. Let $\hat{k}_{o_2}$ be the minimum value such that either $\mathcal{M}^1(\theta_i, \theta_{-i}; (\hat{k}_{o_2}, k_{-o_2})) = o_2$ or $\mathcal{M}^1(\theta'_i, \theta_{-i}; (\hat{k}_{o_2}, k_{-o_2})) = o_2$. At the first such value of $\hat{k}_{o_2}$, only one of these two events will happen because $U^{out}(\theta_i, o_1) - U^{out}(\theta_i, o_2)$ and $U^{out}(\theta'_i, o_1) - U^{out}(\theta'_i, o_2)$ differ by at least 1. Moreover, we have $\hat{k}_{o_2} \leq k_{o_2} + M$ because with $\lambda_{o_2} = k_{o_2}$ we have $V_{o_1} - V_{o_2} \leq M$ for either report θ_i or θ'_i. Since $\Pr[\lambda_{o_2} = k] \propto \exp(-\epsilon \cdot |k|/(M \cdot |O|))$, we have $\Pr[\lambda_{o_2} = k_{o_2}] \leq \exp(\epsilon/|O|) \cdot \Pr[\lambda_{o_2} = \hat{k}_{o_2}]$. Furthermore, there can be at most M such values of k_{o_2}. Thus,

$$\begin{aligned}
&\Pr_r[\lambda_{-o_2} = k_{-o_2} \wedge \mathcal{M}^1(\theta_i, \theta_{-i}; r) = \mathcal{M}^1(\theta'_i, \theta_{-i}; r) = o_1 \wedge \mathcal{M}^2(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^2(\theta'_i, \theta_{-i}; r) \text{ on } o_2] \\
&\leq M e^{\epsilon/|O|} \Pr_r[\lambda_{-o_2} = k_{-o_2} \wedge \mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r) \wedge \mathcal{M}^1(\theta_i, \theta_{-i}; r) \in \{o_1, o_2\} \\
&\wedge \mathcal{M}^1(\theta'_i, \theta_{-i}; r) \in \{o_1, o_2\}]
\end{aligned}$$

Summing over all $o_1 \neq o_2$ and k_{-o_2} gives us the lemma. The factor 2 in the lemma statement is due to the fact that

$$\begin{aligned}
&\sum_{o_1 \neq o_2, k_{o_2}} \Pr_r[\lambda_{-o_2} = k_{-o_2} \wedge \mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r) \wedge \mathcal{M}^1(\theta_i, \theta_{-i}; r) \in \{o_1, o_2\} \wedge \mathcal{M}^1(\theta'_i, \theta_{-i}; r) \in \{o_1, o_2\}] \\
&= 2 \Pr_r[\mathcal{M}^1(\theta_i, \theta_{-i}; r) \neq \mathcal{M}^1(\theta'_i, \theta_{-i}; r)].
\end{aligned}$$

□

Combining Lemmas 7.3 and 7.4, we have

$$\begin{aligned}
\mathrm{SD}(\mathcal{M}(\theta_i, \theta_{-i}), \mathcal{M}(\theta'_i, \theta_{-i})) \leq |O| \cdot (1 + 2Me^{\epsilon/|O|}) \cdot (&\mathrm{E}[U^{out}(\theta_i, \mathcal{M}^1(\theta_i, \theta_{-i})) - P(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}))] \\
&- \mathrm{E}[U^{out}(\theta_i, \mathcal{M}^1(\theta'_i, \theta_{-i})) - P(\theta'_i, \mathcal{M}(\theta_i, \theta_{-i}))]).
\end{aligned}$$

Applying Lemma 5.2 gives us our theorem.

THEOREM 7.5. *Mechanism 7.1 is truthful in expectation and individually rational for player i provided that, for some function F_i:*

(*1*) *Player i's privacy utility U_i^{priv} satisfies Assumption 3.1 with privacy bound function F_i, and*
(*2*) $2F_i(e^{\epsilon}) \cdot |O| \cdot (1 + 2Me^{\epsilon/|O|}) \leq 1$.

In particular, if all players have the same privacy bound function $F_i = F$, it suffices to take ϵ to be a sufficiently small constant depending only on M and $|O|$ (and not the number n of players).

Truthfulness in expectation relies on players being risk neutral in terms of their privacy utility so that it is acceptable that with some low probability, the privacy costs are larger than their utility from the outcome. An alternative approach that does not rely on risk neutrality is to switch from the VCG mechanism to the Expected Externality mechanism. This is a variant on VCG that, rather than charging players the actual externality they impose as in Equation (5), charges them their expected externality

$$E_{\theta \sim p}\left[\sum_{j \neq i} U^{out}(\theta_j, o_{-i}) - U^{out}(\theta_j, o^*)\right], \tag{7}$$

where p is a prior distribution over Θ^n, o_{-i} is the outcome that maximizes the sum of outcome utilities of players other than i, and o^* is the outcome that maximizes the sum of outcome utilities when i is included. Essentially, i is charged the expected amount he would have to pay under VCG given the prior over types. Since the amount players are charged is independent of the actual reports of others, collecting payments has no privacy implications. (The proof of Lemma 7.3 shows that if we only consider the privacy cost of the outcome, then we have universal truthfulness.) However, the use of a prior means that the truthfulness guarantee only holds in a Bayes-Nash equilibrium. On the other hand, this mechanism does have other nice properties such as being adaptable to guarantee budget balance.

Finally, we show that Mechanism 7.1 approximately preserves VCG's efficiency.

Proposition 7.6. *For every profile of types $\theta \in \Theta^n$, if we choose $o \leftarrow \mathcal{M}(\theta)$ using Mechanism 7.1, then we have:*

(1) $\Pr\left[\sum_i U_i^{out}(\theta_i, o) < \max_{o'}\left(\sum_i U_i^{out}(\theta_i, o')\right) - \Delta\right] \le 2|O| \cdot e^{-\epsilon\Delta/(2M\cdot|O|)}$,

(2) $\mathrm{E}\left[\sum_i U_i^{out}(\theta_i, o)\right] \ge \max_{o'}\left(\sum_i U_i^{out}(\theta_i, o')\right) - O(|O|^2 \cdot M/\epsilon)$.

Proof. Let $o^{**} = \mathrm{argmax}_o\, U_j^{out}(\theta_j, o)$. For the output o^* of Mechanism 7.1, we have:

$$\begin{aligned}\sum_j U_j^{out}(\theta_j, o^*) &= V_{o^*} - \lambda_{o^*} - o^*/|O| \\ &\ge V_{o^{**}} - \lambda_{o^*} - o^*/|O| \\ &= \left(\max_o U_j^{out}(\theta_j, o)\right) + \lambda_{o^{**}} + o^{**}/|O| - \lambda_{o^*} - o^*/|O| \\ &> \left(\max_o U_j^{out}(\theta_j, o)\right) - \max_o(\lambda_o - \lambda_{o^{**}}) - 1.\end{aligned}$$

So we are left with bounding $\max_o(\lambda_o - \lambda_{o^{**}})$ for random variables λ_o such that $\Pr[\lambda_o = k] \propto \exp(-\epsilon \cdot |k|/(M \cdot |O|))$. For each o,

$$\Pr[\lambda_o - \lambda_{o^{**}} \ge \Delta] \le \Pr[\lambda_o \ge \Delta/2] + \Pr[\lambda_{o^**} \le -\Delta/2] \le 2\exp(-\epsilon\Delta/(2M \cdot |O|)).$$

Taking a union bound over the choices for o completes the high probability bound. For the expectation, we have:

$$\mathrm{E}[\max_o(\lambda_o - \lambda_{o^{**}})] \le \mathrm{E}\left[\sum_o |\lambda_o|\right] = |O| \cdot O\left(M \cdot |O|/\epsilon\right).$$

□

Thus, the social welfare is within $\tilde{O}(|O|^2) \cdot M/\epsilon$ of optimal, both in expectation and with high probability. Like with Proposition 4.3, these bounds are independent of the number n of participants, so we obtain asymptotically optimal social welfare as $n \to \infty$. Also like the discussion after Proposition 4.3, by taking $\epsilon = \epsilon(n)$ to be such that $\epsilon = o(1)$ and $\epsilon = \omega(1/n)$ (e.g., $\epsilon = 1/\sqrt{n}$), the sum of privacy utilities is also a vanishing fraction of n (for participants satisfying Assumption 3.1 with a common privacy bound function F).

8. A BAYESIAN PERSPECTIVE

We now provide a Bayesian interpretation of our privacy model and discuss several of the model's limitations.

Our modelling of privacy in Section 3 is motivated in part by viewing privacy as a concern about *other's beliefs about you.* Fix a randomized mechanism $\mathcal{M} : \Theta^n \times \mathcal{R} \to O$, a player $i \in [n]$, and a profile $\theta_{-i} \in \Theta^{n-1}$ of other player's reports. Suppose that an adversary has a prior T_i on the type of player i, as well as a prior S_i on the strategy $\sigma : \Theta \to \Theta$ played by player i. Then upon seeing an outcome o from the mechanism, the adversary should replace T_i with a posterior T_i' computed according to Bayes' Rule as follows:

$$\begin{aligned}\Pr[T_i' = \theta_i] &= \Pr[T_i = \theta_i | \mathcal{M}(S_i(T_i), \theta_{-i}) = o] \\ &= \Pr[T_i = \theta_i] \cdot \frac{\Pr[\mathcal{M}(S_i(T_i), \theta_{-i}) = o | T_i = \theta_i]}{\Pr[\mathcal{M}(S_i(T_i), \theta_{-i}) = o]}.\end{aligned}$$

Thus if we set $x = \max_{\theta',\theta''\in\Theta}(\Pr[\mathcal{M}(\theta',\theta_{-i}) = o] / \Pr[\mathcal{M}(\theta'',\theta_{-i}) = o])$ (the argument of F_i in Assumption 3.1), then we have

$$x^{-1} \cdot \Pr[T_i = \theta_i] \leq \Pr[T'_i = \theta_i] \leq x \cdot \Pr[T_i = \theta_i].$$

So if x is close to 1, then the posterior T'_i is close to the prior T_i, having the same probability mass functions within a factor of x, and consequently having statistical difference at most $x - 1$. Thus, Assumption 3.1 can be justified by asserting that "if an adversary's beliefs about player i do not change much, then it has a minimal impact on player i's privacy utility." One way to think of this is that player i has some smooth value function of the adversary's beliefs about her, and her privacy utility is the difference of the value function after and before the Bayesian updating. This reasoning follows the lines of Bayesian interpretations of differential privacy due to Dwork and McSherry, and described in [Kasiviswanathan and Smith 2008].

This Bayesian modelling also explains why we do not include the strategy played by i in the privacy utility function U_i^{priv}. How a Bayesian adversary updates its beliefs about player i based on the outcome do not depend on the actual strategy played by i, but rather on the adversary's beliefs about that strategy, denoted by S_i in the above discussion. Given that our mechanisms are truthful, it is most natural to consider S_i as the truthful strategy (i.e., the identity function). If the Bayesian adversary values possessing a correct belief about players, this is analogous to a notion of equilibrium. If we treat the adversary as another player then if the players report truthfully and the adversary assumes the players report truthfully each is responding optimally to the other. However, if player i can successfully convince the adversary that she will follow some other strategy S_i, then this can be implicitly taken into account in U_i^{priv}. (But if player i further deviates from S_i, this should not be taken into account, since the adversary's beliefs will be updated according to S_i.)

Our modelling of privacy in terms of other's beliefs is subject to several (reasonable) critiques:

- Sometimes a small, continuous change in beliefs can result in discrete choices that have a large impact in someone's life. For example, consider a ranking of potential employees to hire, students to admit, or suitors to marry—a small change in beliefs about a candidate may cause them to drop one place in a ranking, and thereby not get hired, admitted, or married. On the other hand, the candidate typically does not know exactly where such a threshold is and so from their perspective the small change in beliefs could be viewed as causing a small change in the probability of rejection.
- Like in differential privacy, we only consider an adversary's beliefs about player i *given the rest of the database*. (This is implicit in us considering a fixed θ_{-i} in Assumption 3.1.) If an adversary believes that player i's type is correlated with the other players (e.g., given by a joint prior T on Θ^n), then conditioning on $T_{-i} = \theta_{-i}$ may already dramatically change the adversary's beliefs about player i. For example, if the adversary knew that all n voters in a given precinct prefer the same candidate (but don't know which candidate that is), then conditioning on θ_{-i} tells the adversary who player i prefers. We don't measure the (dis)utility for leaking this kind of information. Indeed, the differentially private election mechanism of Theorem 4.2 will leak the preferred candidate in this example (with high probability).
- The word "privacy" is used in many other ways. Instead of being concerned about other's beliefs, one may be concerned about self-representation (e.g., the effect that reporting a given type may have on one's self-image).

Finally, we remark that for some of the mechanism design settings we consider, privacy may not be a major concern on its own. For example, in a deterministic majority-vote election with a large number of voters, each voter's privacy may intuitively be protected because the other voters provide sufficient noise to obscure her vote. However, just like in our mechanism, the limited influence that voters have on the outcome also means that they gain very little in outcome utility from being truthful. Thus, our model and analysis can be viewed as justifying truthfulness in such real-life election scenarios. We also note that introducing noise artificially, as in differentially private mechanisms, has significant advantages over relying on natural noise, particularly with respect to composition. (For example, if we run an election again, with a new voter the second time, and the outcome changes, then the new voter's preference has been revealed despite any uncertainty about the other voter's votes. Adding independent artificial noise to each vote solves this problem, and ensures differential privacy.)

9. MORE GENERAL PRIVACY MODELS

In this section, we introduce more general models of privacy in mechanism design, and explain analogues of our results for these models. The purpose of introducing these models is to elucidate assumptions that were implicit in our earlier modelling, explore the extent to which we can allow players' privacy losses to depend on their strategies, and to discuss whether the revelation principle holds in our setting.

Here we consider non-direct-revelation mechanisms, where each player chooses an *action* from an action space X, and the mechanism is a (possibly randomized) mapping $\mathcal{M} : X^n \to O$ from actions to outcomes. Each player i chooses her action using a (possibly randomized) *strategy* $\sigma_i : \Theta \to X$. We let $\mathcal{S}$ denote the set of all randomized strategies that players can use.

In the greatest generality, we could allow player i's privacy utility U_i^{priv} to depend on a profile θ of all of the players' types, the outcome o, the mechanism $\mathcal{M}$, a profile σ of all of the players' strategies, and a deviation σ_i' taken by player i:

$$U_i^{priv} : \Theta^n \times O \times \{\mathcal{M} : X^n \times \mathcal{R} \to O\} \times \mathcal{S}^n \times \mathcal{S} \to \mathbb{R}. \tag{8}$$

The most important differences with Equation (3) is that we are allowing the privacy loss for player i, $U_i^{priv}(\theta, o, \mathcal{M}, \sigma, \sigma_i')$ to depend on full randomized strategies σ_i of all n players, whereas in Equation (3), we only allow it to depend on the specific *reports* $\theta_i' = \sigma_i(\theta_i)$ of players other than player i. (The dependence on the actual types $\theta = (\theta_1, \ldots, \theta_n)$ is not so significant since our dominant-strategy solution concept quantifies over all player types, and we allow player-specific utility functions U_i^{priv}.) The reason for keeping both an initial strategy σ_i for player i and a deviation σ_i' comes from the Bayesian perspective of Section 8; a Bayesian adversary might believe that player i is playing σ_i when she actually deviates and plays σ_i'.

Since U_i^{priv} has in its arguments enough information to compute the entire output distribution of the mechanism, i.e. $\mathcal{M}(\sigma_1(\theta_1), \sigma_2(\theta_2), \ldots, \sigma_n(\theta_n))$, it also makes sense to omit the outcome and allow for a direct definition of an "expected" privacy utility (which could be obtained by taking an expectation over outcomes, or allow for more general models):

$$EU_i^{priv} : \Theta^n \times \{\mathcal{M} : X^n \times \mathcal{R} \to O\} \times \mathcal{S}^n \times \mathcal{S} \to \mathbb{R}. \tag{9}$$

With these models of privacy utility functions, we obtain the following solution concepts in analogy to Definition 3.4:

Definition 9.1 (*truthfulness with privacy, generalized*). Consider a mechanism design problem with n players, type space Θ, action space X, and outcome space O. Let the *expected utility* function of player i be $EU_i = EU_i^{out} + EU_i^{priv}$, where $EU_i^{out}(\theta, \mathcal{M}, \sigma, \sigma_i') = \mathrm{E}[U_i^{out}(\theta_i, \mathcal{M}(\sigma_i'(\theta_i), \sigma_{-i}(\theta_{-i})))]$ and EU_i^{priv} has the syntax of Equation (9).

A strategy profile $\sigma \in \mathcal{S}^n$ is *dominant in expectation for player i* with type θ_i if for all alternative strategies σ_i' and all type profiles $\theta_{-i} \in \Theta^{n-1}$, we have

$$EU_i(\theta, \mathcal{M}, \sigma, \sigma_i) \geq EU_i(\theta, \mathcal{M}, \sigma, \sigma_i').$$

Let the *per-outcome utility* function of player i be $U_i = U_i^{out} + U_i^{priv}$, where U_i^{priv} is as in Equation (8). We say that a strategy profile $\sigma \in \mathcal{S}^n$ is *universally dominant for player i* with type θ_i if for all alternative strategies σ_i', all type profiles $\theta_{-i} \in \Theta^{n-1}$, and all values of $r \in \mathcal{R}$, we have:

$$\mathrm{E}[U_i(\theta, \mathcal{M}(\sigma_i(\theta_i), \sigma_{-i}(\theta_{-i}); r), \mathcal{M}, \sigma, \sigma_i)] \geq \mathrm{E}[U_i(\theta, \mathcal{M}(\sigma_i'(\theta_i), \sigma_{-i}(\theta_{-i}); r), \mathcal{M}, \sigma, \sigma_i')],$$

where the expectations are taken over the randomized strategies in σ and σ'_i.

We say that $\mathcal{M}$ is *truthful in expectation (resp., universally truthful) for player i* if $X = \Theta$ and the identity profile is a dominant in expectation (resp., universally dominant) strategy for player i.

With these definitions, we have the following generalization of Assumption 3.1:

Assumption 9.2 (generalized privacy-value assumption).

$$\forall \theta \in \Theta^n, o \in O, \mathcal{M}, \sigma \in \mathcal{S}^n, \sigma'_i \in \mathcal{S} : \left|U_i^{priv}(\theta, o, \mathcal{M}, \sigma, \sigma'_i)\right| \le F_i\left(\max_{\sigma''_i,\sigma'''_i \in \mathcal{S}} \frac{\Pr\left[\mathcal{M}(\sigma''_i(\theta_i), \sigma_{-i}(\theta_{-i})) = o\right]}{\Pr\left[\mathcal{M}(\sigma'''_i(\theta_i), \sigma_{-i}(\theta_{-i})) = o\right]}\right)$$

where $F_i : [1, \infty) \to [0, \infty]$ *is a* privacy-bound *function with the property that* $F_i(x) \to 0$ *as* $x \to 1$, *and the probabilities are taken over the randomness of both* $\mathcal{M}$ *and the strategies* $\sigma''_i, \sigma'''_i, \sigma_{-i}$.

To make use of it, we will again use differential privacy, which we restate using this section's more general notation.

Definition 9.3. A mechanism $\mathcal{M} : X^n \times \mathcal{R} \to O$ is *ϵ-differentially private* iff

$$\forall x_{-i} \in X^n, o \in O, \qquad \max_{x'_i, x''_i \in \mathcal{S}} \frac{\Pr\left[\mathcal{M}(x'_i, x_{-i}) = o\right]}{\Pr\left[\mathcal{M}(x''_i, x_{-i}) = o\right]} \le e^{\epsilon}.$$

By inspection of Assumption (9.2) and the definition of differential privacy, we have the following result.

Proposition 9.4. *If* $\mathcal{M}$ *is ϵ-differentially private, then for all players i whose utility functions satisfy Assumption (3.1), all* $\theta \in \Theta^n, o \in O, \sigma \in \mathcal{S}^n, \sigma'_i \in \mathcal{S}$ *we have* $\left|U_i^{priv}(\theta, o, \mathcal{M}, \sigma, \sigma'_i)\right| \le F_i(e^{\epsilon})$.

Under this assumption and the further requirement that a player's privacy utility does not depend on the strategy σ'_i that she actually follows, we obtain the following generalized analysis of the differentially private two-candidate voting mechanism (Theorem 4.2):

Theorem 9.5. *Mechanism 4.1 is universally truthful for player i provided that, for some function* F_i*:*

(1) *Player i's privacy utility* U_i^{priv} *satisfies Assumption 9.2 with privacy bound function* F_i,
(2) *Player i's privacy utility* U_i^{priv} *does not depend on the strategy* σ'_i *followed by player i, and*
(3) $U_i^{out}(\theta_i, \theta_i) - U_i^{out}(\theta_i, \neg\theta_i) \ge 2F_i(e^{\epsilon})$,

A justification for the assumption that U_i^{priv} does not directly depend on the strategy σ'_i actually followed is that σ'_i is not visible to an external observer (such as a privacy adversary) except through its effect on the outcome o.

Proof. Fix the actual type $\theta_i \in \{A, B\}$ of player i, a type profile θ_{-i} and strategy profile σ_{-i} of the other players, and a choice r for $\mathcal{M}$'s randomness. The truthful strategy is $\sigma(\theta) = \theta$ and an alternate strategy for player i is now σ'_i, which consists of two probability distributions on $\{A, B\}$ (one for each type).

We need to show that

$$\mathrm{E}[U_i(\theta, \mathcal{M}(\theta_i, \theta_{-i}; r), \mathcal{M}, \sigma, \sigma_i)] \ge \mathrm{E}[U_i(\theta, \mathcal{M}(\sigma'_i(\theta_i), \theta_{-i}; r), \mathcal{M}, \sigma, \sigma'_i)],$$

where the expectations are taken over the (potentially) randomized strategy σ'_i. Equivalently:

$$\begin{aligned} &\mathrm{E}[U_i^{out}(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r)) + U_i^{priv}(\theta, \mathcal{M}(\theta_i, \theta_{-i}; r), \mathcal{M}, \sigma)] \qquad (10)\\ &\ge \mathrm{E}[U_i^{out}(\theta_i, \mathcal{M}(\sigma'_i(\theta_i), \theta_{-i}; r)) + U_i^{priv}(\theta, \mathcal{M}(\sigma'_i(\theta_i), \theta_{-i}; r), \mathcal{M}, \sigma)], \end{aligned}$$

where again the expectations are taken over the randomized strategy σ'_i and we have omitted the final argument of U_i^{priv} (set to σ_i and σ'_i, respectively) because by assumption it is irrelevant. Since $\sigma'_i(\theta_i)$ is simply a distribution over pure strategies θ'_i, by averaging it suffices to show that for every θ'_i, we have:

$$\begin{aligned} &U_i^{out}(\theta_i, \mathcal{M}(\theta_i, \theta_{-i}; r)) + U_i^{priv}(\theta, \mathcal{M}(\theta_i, \theta_{-i}; r), \mathcal{M}, \sigma) \\ &\geq U_i^{out}(\theta_i, \mathcal{M}(\theta'_i, \theta_{-i}; r)) + U_i^{priv}(\theta, \mathcal{M}(\theta'_i, \theta_{-i}; r), \mathcal{M}, \sigma), \end{aligned} \tag{11}$$

From this point, the proof exactly follows that of Theorem 4.2. Specifically, setting $o = \mathcal{M}(\theta_i, \theta_{-i}; r)$ and $o' = \mathcal{M}(\theta'_i, \theta_{-i}; r)$ and rearranging terms, we need to show:

$$U_i^{out}(\theta_i, o) - U_i^{out}(\theta_i, o') \geq U_i^{priv}(\theta, o', \mathcal{M}, \sigma) - U_i^{priv}(\theta, o, \mathcal{M}, \sigma). \tag{12}$$

We consider two cases:

Case 1: $o = o'$. In this case, Inequality (12) holds because both the left-hand and right-hand sides are zero.

Case 2: $o \neq o'$. This implies that $o = \theta_i$ and $o' = \neg\theta_i$. (If player i's report has any effect on the outcome of the differentially private voting mechanism, then it must be that the outcome equals player i's report.) Thus the left-hand side of Inequality (12) equals $U_i^{out}(\theta_i, \theta_i) - U_i^{out}(\theta_i, \neg\theta_i)$. By Proposition 9.4, the right-hand side of Inequality (12) is at most $2F(e^\epsilon)$. Thus, Inequality (12) holds by hypothesis.

□

When using expected utility, we can allow a player's privacy utility to depend on her strategy, but the effect of changing a strategy should be bounded by the change it causes in the output distribution of the mechanism. That is:

Assumption 9.6 (influence-bounded privacy).

$$\forall \theta \in \Theta^n, \sigma \in \mathcal{S}^n, \sigma'_i \in \mathcal{S} : \left| EU_i^{priv}(\theta, \mathcal{M}, \sigma, \sigma_i) - EU_i^{priv}(\theta, \mathcal{M}, \sigma, \sigma'_i) \right| \leq \epsilon_i \cdot SD(\mathcal{M}(\sigma(\theta)), \mathcal{M}(\sigma'_i(\theta_i), \sigma_{-i}(\theta_{-i}))).$$

If this condition holds, we say that $\mathcal{M}$ provides influence-bounded privacy ϵ_i *to player i.*

If a player's privacy utility does not depend on her strategy, then influence-bounded privacy follows from Assumption 9.2 and the mechanism $\mathcal{M}$ being differentially private (but influence-bounded privacy is more general, in that it allows a dependence on a player's strategy):

Lemma 9.7. *Consider a mechanism design problem with n players, type space Θ, action space X, and outcome space O. Let player i have type $\theta_i \in \Theta_i$ and a utility function $U_i = U_i^{out} + U_i^{priv}$ with the syntax of Equation (8). Define $EU_i^{priv}(\theta, \mathcal{M}, \sigma, \sigma'_i) = \mathrm{E}[U_i^{priv}(\theta, \mathcal{M}(\sigma'_i(\theta_i), \sigma_{-i}(\theta_{-i})), \mathcal{M}, \sigma, \sigma'_i)]$. Suppose:*

(1) Randomized mechanism $\mathcal{M} : \Theta^n \times \mathcal{R} \to O$ is ϵ-differentially private,
(2) U_i^{priv} satisfies Assumption (9.2) with privacy bound F_i, and
(3) U_i^{priv} does not depend on the strategy σ'_i of player i.

Then, $\mathcal{M}$ provides $2F_i(e^\epsilon)$ influence-bounded privacy to player i.

The proof of this lemma is essentially identical to that of Lemma 5.2.

Proof. For every two discrete random variables X and Y taking values in a universe $\mathcal{U}$, and every function $f : \mathcal{U} \to [-1, 1]$, it holds that $|\,\mathrm{E}[f(X)] - \mathrm{E}[f(Y)]| \leq 2\mathrm{SD}(X, Y)$. (The f that maximizes the left-hand side sets $f(x) = 1$ when $\Pr[X = x] > \Pr[Y = x]$ and sets $f(x) = -1$ otherwise.) Take $\mathcal{U} = O$, $X = \mathcal{M}(\sigma_i(\theta_i), \sigma_{-i}(\theta_{-i}))$, $Y = \mathcal{M}(\sigma'_i(\theta_i), \sigma_i(\theta_{-i}))$, and $f(o) = U_i^{priv}(\theta, o, \mathcal{M}, \sigma)/F_i(e^\epsilon)$. (Again we omit the last argument of U_i^{priv} because by assumption it is irrelevant.) By Proposition 9.4, we have $f(o) \in [-1, 1]$. □

Now we can generalize the analysis of the voting mechanism to obtain truthfulness in expectation even when privacy utilities depend on players' strategies, provided that they are influence-bounded according to Assumption 9.6.

Theorem 9.8. *Mechanism 4.1 is truthful in expectation for player i assuming that it provides influence-bounded privacy at most $U_i^{out}(\theta_i, \theta_i) - U_i^{out}(\theta_i, \neg\theta_i)$ to player i.*

Proof. We need to show that

$$EU_i(\theta, \mathcal{M}, \sigma, \sigma_i) \geq EU_i(\theta, \mathcal{M}, \sigma, \sigma'_i),$$

or equivalently

$$\begin{aligned} &\mathrm{E}[U_i^{out}(\theta_i, \mathcal{M}(\sigma_i(\theta_i), \sigma_{-i}(\theta_{-i})))] - \mathrm{E}[U_i^{out}(\theta_i, \mathcal{M}(\sigma'_i(\theta_i), \sigma_{-i}(\theta_{-i})))] \\ &\geq EU_i^{priv}(\theta, \mathcal{M}, \sigma, \sigma_i) - EU_i^{priv}(\theta, \mathcal{M}, \sigma, \sigma'_i)]. \end{aligned} \tag{13}$$

By assumption, the right hand side of Inequality 13 is at most $(U_i^{out}(\theta_i, \theta_i) - U_i^{out}(\theta_i, \neg\theta_i)) \cdot \mathrm{SD}(\mathcal{M}(\sigma(\theta)), \mathcal{M}(\sigma'_i(\theta_i), \sigma_{-i}(\theta_{-i})))$, which by definition is exactly the value of the left hand side. □

The Revelation Principle.. We now have a definitional framework that allows us to discuss the extent to which the revelation principle applies in our setting. Recall that the revelation principle says that if we have a non-direct-revelation mechanism $\mathcal{M}$ with a dominant-strategy equilibrium profile σ, then we can simulate it by a direct revelation mechanism $\overline{\mathcal{M}}$ where $\overline{\mathcal{M}}(\theta) = \mathcal{M}(\sigma(\theta))$. That is, the new mechanism $\overline{\mathcal{M}}$ takes the (reported) types of the players, evaluates their strategies σ for them, and then runs the original mechanism. Previous papers (particularly, Ghosh and Roth [2011]) have argued that the revelation principle should not hold in contexts where agents value privacy, because it requires revealing more information. However, we (and previous papers such as [Ghosh and Roth 2011]) are working in a model where the mechanism is trusted and the only privacy concerns being modelled are with respect to an external adversary that views the outcome (which is the same under $\overline{\mathcal{M}}$ as under $\mathcal{M}$). In such a case, a revelation principle intuitively should hold.

However, it is subtle to formalize this intuition, since the modelling of agents that value privacy requires that the privacy-utility functions depend on the mechanism $\mathcal{M}$ itself. If we change the mechanism to $\overline{\mathcal{M}} = \mathcal{M} \circ \sigma$, the utility functions can change dramatically, and we can no longer relate strategic considerations in the two mechanisms. Thus, we need to impose more structure on the way in which the utilities depend on $\mathcal{M}$ (and σ). Taking the Bayesian perspective of Section 8, it is natural to require that the privacy-utility for player i when deviating from equilibrium σ by σ'_i depends only on the functions $\mathcal{M} \circ \sigma$ and $\mathcal{M} \circ (\sigma'_i, \sigma_{-i})$, along with the type profile θ and the outcome o. The function $\mathcal{M} \circ \sigma$ reflects the mapping from types to outcome that a Bayesian adversary expects the players to play and the function $\mathcal{M} \circ (\sigma'_i, \sigma_{-i})$ is the mapping induced by player i's deviation. Implicit in this modelling is the view that an adversary is not colluding with any of the other players, and in particular cannot see their randomization or report. Under this modelling, the revelation principle does hold:

Theorem 9.9 (Revelation Principle with Privacy). *Consider a mechanism design problem with n players, type space Θ, action space X, and outcome space O. Let the* expected utility *function of player i be $EU_i = EU_i^{out} + EU_i^{priv}$, where $EU_i^{out}(\theta, \mathcal{M}, \sigma, \sigma'_i) = \mathrm{E}[U_i^{out}(\theta_i, \mathcal{M}(\sigma'_i(\theta_i), \sigma_{-i}(\theta_{-i})))]$ and EU_i^{priv} has the syntax of Equation (9) and satisfies $EU_i^{priv}(\theta, \mathcal{M}, \sigma, \sigma'_i) = EV_i^{priv}(\theta, \mathcal{M} \circ \sigma, \mathcal{M} \circ (\sigma'_i, \sigma_{-i}))$ for some function EV_i^{priv}.*

Then if σ_i is dominant in expectation for player i in mechanism $\mathcal{M}$, it follows that the direct revelation mechanism $\overline{\mathcal{M}} = \mathcal{M} \circ \sigma$ is truthful in expectation for player i. Moreover, for every type *profile θ, the players' utilities and the joint distribution on outcomes are the same under $\overline{\mathcal{M}}$ with the truthful strategy profile $\overline{\sigma}$ as they are under $\mathcal{M}$ with σ.*

Proof. We begin with the second claim. Let $\overline{\sigma}$ be the truthful strategy profile for $\overline{\mathcal{M}}$. Given a mechanism $\mathcal{M}$ and strategy profile σ, for all θ the distribution over outcomes is the law of $\mathcal{M}(\sigma(\theta)) = \mathcal{M} \circ \sigma(\theta) = \overline{\mathcal{M}}(\overline{\sigma}(\theta))$. Thus they have the same distribution over outcomes. Similarly the distribution over outcome utilities is the law of $U_i^{out}(\theta_i, M(\sigma(\theta)))$ in each case and so the joint distribution over outcomes and outcome utility is the same for each mechanism. This implies that $EU_i^{out}(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}_i) = EU_i^{out}(\theta, \mathcal{M}, \sigma, \sigma_i)$. We also have

$$\begin{aligned} EU_i^{priv}(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}_i) &= EV_i^{priv}(\theta, \overline{\mathcal{M}} \circ \overline{\sigma}, \overline{\mathcal{M}} \circ (\overline{\sigma}_i, \overline{\sigma}_{-i})) \\ &= EV_i^{priv}(\theta, \mathcal{M} \circ \sigma, \mathcal{M} \circ (\sigma_i, \sigma_{-i})) \\ &= EU_i^{priv}(\theta, \mathcal{M}, \sigma, \sigma_i) \end{aligned}$$

. Thus the utilities are the same under $\overline{\mathcal{M}}$ with $\overline{\sigma}$ as under $\mathcal{M}$ with σ.

Now we prove the first claim, about truthfulness in expectation. Given an arbitrary strategy $\overline{\sigma}'_i : \Theta \to \Delta(\Theta)$, we need to show that

$$EU_i(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}_i) \geq EU_i(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}'_i),$$

Let $\sigma'_i(\theta) = \sigma_i(\overline{\sigma}'_i(\theta))$. Then $\overline{\mathcal{M}}(\overline{\sigma}'_i(\theta), \overline{\sigma}_{-i}(\theta_{-i})) = \mathcal{M}(\sigma'_i(\theta), \sigma_{-i}(\theta_{-i}))$. Thus, by the claim regarding utilities proven above and the fact that $\mathcal{M}$ is dominant in expectation, we have:

$$\begin{aligned} EU_i(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}_i) &= EU_i(\theta, \mathcal{M}, \sigma, \sigma_i) \\ &\geq EU_i(\theta, \mathcal{M}, \sigma, \sigma'_i) \\ &= \mathrm{E}[U_i^{out}(\theta_i, \mathcal{M}(\sigma'_i(\theta_i), \sigma_{-i}(\theta_{-i})))] + EV_i^{priv}(\theta, \mathcal{M} \circ \sigma, \mathcal{M} \circ (\sigma'_i, \sigma_{-i})) \\ &= \mathrm{E}[U_i^{out}(\theta_i, \overline{\mathcal{M}}(\overline{\sigma}'_i(\theta_i), \sigma_{-i}(\theta_{-i})))] + EV_i^{priv}(\theta, \overline{\mathcal{M}} \circ \overline{\sigma}, \overline{\mathcal{M}} \circ (\overline{\sigma}'_i, \overline{\sigma}_{-i})) \\ &= EU_i^{out}(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}'_i) + EU_i^{priv}(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}'_i) \\ &= EU_i(\theta, \overline{\mathcal{M}}, \overline{\sigma}, \overline{\sigma}'_i). \end{aligned}$$

□

We note that the conditions on EU_i^{priv} in Theorem 9.9 hold in particular under the Bayesian adversary model of Section 8 (which is consistent with, and indeed motivated, the model used in the main results of the paper). That is, suppose a player's privacy-utility is the function of a Bayesian adversary's posterior belief T'_i of their type. Fixing the adversary's joint prior T on all the players types, given an outcome o, the posterior T'_i is computed as:

$$\Pr[T'_i = \theta'_i] = \Pr[T_i = \theta'_i | \mathcal{M}(\sigma(T)) = o].$$

(In Section 8, we took the other players' types as fixed at θ_{-i} in order to apply differential privacy, but the revelation principle holds with an arbitrary joint prior.) Thus, the per-outcome privacy utility U_i^{priv} depends on o, and $\mathcal{M} \circ \sigma$ (but not any other property of $\mathcal{M}$ or σ). Now, when player i deviates to strategy σ'_i, the expected privacy utility EU_i^{priv} is obtained by taking the expectation of U_i^{priv} over $o \leftarrow \mathcal{M}(\sigma'_i(\theta_i), \sigma_{-i}(\theta_{-i}))$. Thus player i's expected privacy utility EU_i^{priv} is a function of only θ, $\mathcal{M} \circ \sigma$, and $\mathcal{M} \circ (\sigma'_i, \sigma_{-i})$, as needed for the revelation principle.

Another example of privacy-utility functions that meet the conditions of Theorem 9.9 is a mutual information measure used by Xiao [2011]: $EU_i^{priv} = I(T_i; \mathcal{M}(\sigma'_i(T_i), \sigma_{-i}(T_{-i})))$, where again $T = (T_i, T_{-i})$ is a joint prior on the types of all players. (Xiao's notion of privacy is discussed more in Appendix A.)

An alternative justification for the restriction to direct-revelation mechanisms in our paper is that the properties we require are preserved under passing from $\mathcal{M}$ to $\overline{\mathcal{M}} = \mathcal{M} \circ \sigma$. Intuitively, all we use is that players have a limited influence on the output distribution (in order to bound their privacy-utility functions), and their influence only decreases as we move from $\mathcal{M}$ to $\overline{\mathcal{M}}$. In particular, if $\mathcal{M}$ is ϵ-differentially private, then so is $\overline{\mathcal{M}}$. Moreover, for every outcome o and $\theta \in \Theta^n$, we have:

$$\max_{\theta'_i, \theta''_i \in \mathcal{S}} \frac{\Pr\left[\overline{\mathcal{M}}(\theta'_i, \theta_{-i}) = o\right]}{\Pr\left[\overline{\mathcal{M}}(\theta''_i), \theta_{-i}) = o\right]} \leq \max_{\sigma''_i, \sigma'''_i \in \mathcal{S}} \frac{\Pr\left[\mathcal{M}(\sigma''_i(\theta_i), \sigma_{-i}(\theta_{-i})) = o\right]}{\Pr\left[\mathcal{M}(\sigma'''_i(\theta_i), \sigma_{-i}(\theta_{-i})) = o\right]}$$

Thus if we take Assumptions 3.1 and 9.2, we obtain a stronger upper bound on the privacy utility functions for $\overline{\mathcal{M}}$ than for $\mathcal{M}$, and hence a truthfulness analysis that only uses such an upper bound will work as well for $\overline{\mathcal{M}}$ as for $\mathcal{M}$.

Acknowledgments

This work was inspired by discussions under the Harvard Center Research for Computation and Society's "Data Marketplace" project. We are grateful to the other participants in those meetings, including Scott Kominers, David Parkes, Felix Fischer, Ariel Procaccia, Aaron Roth, Latanya Sweeney, and Jon Ullman. We also thank Moshe Babaioff, Kobbi Nissim, Dave Xiao, and the anonymous reviewers for helpful comments and discussions.

A. COMPARISON TO XIAO'S PRIVACY MEASURE[4]

Xiao [2011] measures privacy cost as being proportional to the mutual information between a player's type and the outcome of the mechanism, where the *mutual information* between two jointly distributed random variables X and Y is defined to be

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = \mathop{\mathrm{E}}_{(x,y)\sim(X,Y)}\left[\log \frac{\Pr[(X,Y) = (x,y)]}{\Pr[X = x] \cdot \Pr[Y = y]}\right],$$

where $H(Z) = \mathrm{E}_{z\sim Z}[\log(1/\Pr[Z = z])]$ is Shannon entropy. In order for the mutual information to make sense, Xiao assumes a prior T_i on a player's type and the privacy cost also depends on the strategy $\sigma'_i : \Theta \to \Theta$ played by player i. Accordingly his measure of outcome utility also takes an expectation over the same prior T_i. As mentioned earlier, Xiao's modelling is not a special case of our main definitions, because his modelling of privacy depends on the actual strategy σ_i followed by player i. Nevertheless we can fit it into the more general framework of Section 9. Specifically he takes:

Definition A.1. Let Θ be a type space, O an outcome space, $\mathcal{M} : \Theta^n \times \mathcal{R} \to O$ a randomized direct-revelation mechanism. For a prior T_i on player i's type and $\nu_i \geq 0$ a measure of player i's value for privacy, *Xiao's privacy-utility function* is

$$EU_i(\theta, \mathcal{M}, \sigma, \sigma'_i) = -\nu_i \cdot I(T_i; \mathcal{M}(\sigma'_i(T_i), \sigma_{-i}(\theta_{-i}))).$$

$\mathcal{M}$ is *Xiao-truthful* for player i if it is truthful in expectation for player i in the sense of Definition 9.1.

While mutual information is a natural first choice for measuring privacy, it has several disadvantages compared to our modelling:

— It treats all bits of information the same, whereas clearly one may have different concerns for different aspects of one's private type. For example, one may be a lot more sensitive about the high-order bits of one's salary than the low-order bits.

[4] Subsequent to our work, Xiao has revised his model to use a different, prior-free measure of privacy. This appendix provides a comparison to his original formulation.

— It forces us to consider a prior on a player's type and take expected utility over that prior. Contrast this with the Bayesian interpretation of our privacy modelling described in Section 8. There the prior T_i is only an adversary's beliefs about player i's type, which may be completely incorrect. Player i's utility is computed with respect to his fixed, actual type θ_i.

Nevertheless, we can show that truthfulness with respect to our definitions implies truthfulness with respect to his:

THEOREM A.2. *If $\mathcal{M}$ is truthful in expectation for player i with respect to the privacy utility function*

$$U_i^{priv}(\theta_i, o, \mathcal{M}, \theta_{-i}) = -\nu_i \cdot \log \frac{\Pr[\mathcal{M}(\theta_i, \theta_{-i}) = o]}{\Pr[\mathcal{M}(T_i, \theta_{-i}) = o]},$$

then $\mathcal{M}$ is Xiao-truthful for player i with prior T_i.

We note that the privacy utility function in Theorem A.2 satisfies Assumption 3.1 with $F_i(x) = \nu_i \cdot \log(x)$, and hence all of our truthful mechanisms are also Xiao-truthful.

PROOF. First note that, by Bayes' Rule,

$$U_i^{priv}(\theta_i, o, \mathcal{M}, \theta_{-i}) = -\nu_i \cdot \log \frac{\Pr[\mathcal{M}(T_i, \theta_{-i}) = o | T_i = \theta_i]}{\Pr[\mathcal{M}(T_i, \theta_{-i}) = o]} = -\nu_i \cdot \log \frac{\Pr[(T_i, \mathcal{M}(T_i, \theta_{-i})) = (\theta_i, o)]}{\Pr[T_i = \theta_i] \cdot \Pr[\mathcal{M}(T_i, \theta_{-i}) = o]}. \tag{14}$$

Thus,

$$-\nu_i \cdot I(T_i; \mathcal{M}(T_i, \theta_{-i})) = \mathrm{E}\left[U_i^{priv}(T_i, \mathcal{M}(T_i, \theta_{-i}), \mathcal{M}, \theta_{-i})\right]. \tag{15}$$

To relate the mutual information under strategy σ_i' to U_i^{priv}, we use the notion of *KL divergence* between two random variables X and Y, which is defined as

$$KL(X \| Y) = \mathop{\mathrm{E}}_{x \sim X}\left[\log \frac{\Pr[X = x]}{\Pr[Y = y]}\right].$$

We will use the fact that for a random variable W jointly distributed with X and Y, we have $KL(W, X \| W, Y) \geq KL(X \| Y)$. (This follows from the Log-Sum Inequality [Cover and Thomas 1991].) Taking $W = T_i$, $X = \mathcal{M}(\sigma_i'(T_i), \theta_{-i})$, and $Y = \mathcal{M}(T_i, \theta_{-i})$, we have

$$\begin{aligned} & I(T_i; \mathcal{M}(\sigma_i'(T_i), \theta_{-i})) \\ \geq\ & I(T_i; \mathcal{M}(\sigma_i'(T_i), \theta_{-i})) - KL(T_i, \mathcal{M}(\sigma_i'(T_i)) \| T_i, \mathcal{M}(T_i)) + KL(\mathcal{M}(\sigma_i'(T_i)) \| \mathcal{M}(T_i)) \\ =\ & \mathop{\mathrm{E}}_{(\theta_i, o) \sim (T_i, \mathcal{M}(\sigma_i'(T_i), \theta_{-i}))}\left[\log \frac{\Pr[(T_i, \mathcal{M}(T_i, \theta_{-i})) = (\theta_i, o)]}{\Pr[T_i = \theta_i] \cdot \Pr[\mathcal{M}(T_i, \theta_{-i}) = o]}\right]. \end{aligned}$$

Combining this with Equation (14), we have:

$$-\nu_i \cdot I(T_i; \mathcal{M}(\sigma_i'(T_i), \theta_{-i})) \leq \mathrm{E}\left[U_i^{priv}(T_i, o, \mathcal{M}(\sigma_i'(T_i), \theta_{-i})\right]. \tag{16}$$

By truthfulness in expectation with respect to U_i^{priv}, we have

$$\begin{aligned} & \mathrm{E}[U^{out}(T_i, \mathcal{M}(T_i, \theta_{-i}))] + \mathrm{E}\left[U_i^{priv}(T_i, \mathcal{M}(T_i, \theta_{-i}), \mathcal{M}, \theta_{-i})\right] \\ \geq\ & \mathrm{E}[U^{out}(T_i, \mathcal{M}(\sigma_i'(T_i), \theta_{-i}))] + \mathrm{E}\left[U_i^{priv}(T_i, o, \mathcal{M}(\sigma_i'(T_i), \theta_{-i}, \theta_{-i})\right] \end{aligned} \tag{17}$$

Combining Inequalities (15), (16), and (17) completes the proof. □

REFERENCES

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. 2005. Practical privacy: the SuLQ framework. In *PODS*, Chen Li (Ed.). ACM, 128–138.

Felix Brandt and Tuomas Sandholm. 2008. On the Existence of Unconditionally Privacy-Preserving Auction Protocols. *ACM Trans. Inf. Syst. Secur.* 11, Article 6 (May 2008), 21 pages. Issue 2. DOI:http://dx.doi.org/10.1145/1330332.1330338

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory* (2nd ed.). John Wiley & Sons, Inc.

Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *PODS*. ACM, 202–210.

Yevgeniy Dodis, Shai Halevi, and Tal Rabin. 2000. A Cryptographic Solution to a Game Theoretic Problem. In *CRYPTO (Lecture Notes in Computer Science)*, Mihir Bellare (Ed.), Vol. 1880. Springer, 112–130.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC (Lecture Notes in Computer Science)*, Shai Halevi and Tal Rabin (Eds.), Vol. 3876. Springer, 265–284.

Cynthia Dwork and Kobbi Nissim. 2004. Privacy-Preserving Datamining on Vertically Partitioned Databases. In *CRYPTO (Lecture Notes in Computer Science)*, Matthew K. Franklin (Ed.), Vol. 3152. Springer, 528–544.

Cynthia Dwork and Aaron Roth. 2014. *The Algorithmic Foundations of Differential Privacy*. Now Publishers.

Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. 2010. Boosting and Differential Privacy. In *FOCS*. IEEE Computer Society, 51–60.

Joan Feigenbaum, Aaron D. Jaggard, and Michael Schapira. 2010. Approximate privacy: foundations and quantification (extended abstract). In *ACM Conference on Electronic Commerce*, David C. Parkes, Chrysanthos Dellarocas, and Moshe Tennenholtz (Eds.). ACM, 167–178.

Arpita Ghosh and Aaron Roth. 2011. Selling privacy at auction. In *Proceedings of the 12th ACM conference on Electronic commerce (EC '11)*. ACM, New York, NY, USA, 199–208. DOI:http://dx.doi.org/10.1145/1993574.1993605

Ronen Gradwohl. 2012. Privacy in Implementation. Working Paper. (November 2012).

Ronen Gradwohl and Rann Smorodinsky. 2014. Subjective Perception Games and Privacy. arXiv:1409l.1487. (September 2014).

Zhiyi Huang and Sampath Kannan. 2012. The Exponential Mechanism for Social Welfare: Private, Truthful, and Nearly Optimal. In *FOCS*.

Sergei Izmalkov, Silvio Micali, and Matt Lepinski. 2005. Rational Secure Computation and Ideal Mechanism Design. In *FOCS*. IEEE Computer Society, 585–595.

Shiva Prasad Kasiviswanathan and Adam Smith. 2008. A Note on Differential Privacy: Defining Resistance to Arbitrary Side Information. *CoRR* abs/0803.3946 (2008).

Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy. In *FOCS*. IEEE Computer Society, 94–103.

Moni Naor, Benny Pinkas, and Reuban Sumner. 1999. Privacy preserving auctions and mechanism design. In *ACM Conference on Electronic Commerce*. 129–139.

Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. 2011. Privacy-Aware Mechanism Design. arXiv:1111.3350v1. (November 2011). To Appear in *EC 2012*.

Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. 2010. Approximately Optimal Mechanism Design via Differential Privacy. *CoRR* abs/1004.2888 (2010). To appear in *ITCS 2012*.

Mallesh M. Pai and Aaron Roth. 2013. Privacy and Mechanism Design. *ACM SIGecom Exchangesi* 12 (June 2013), 8–29. Issue 1.

David C. Parkes, Michael O. Rabin, Stuart M. Shieber, and Christopher Thorpe. 2008. Practical secrecy-preserving, verifiably correct and trustworthy auctions. *Electronic Commerce Research and Applications* 7, 3 (2008), 294–312.

David Xiao. 2011. *Is privacy compatible with truthfulness?* Technical Report 2011/005. Cryptology ePrint Archive.

David Xiao. 2013. Is privacy compatible with truthfulness?. In *ITCS*, Robert D. Kleinberg (Ed.). ACM, 67–86.