10

Differential Privacy and Mechanism Design

One of the most fascinating areas of game theory is mechanism design, which is the science of designing incentives to get people to do what you want them to do. Differential privacy has proven to have interesting connections to mechanism design in a couple of unexpected ways. It provides a tool to quantify and control privacy loss, which is important if the people the mechanism designer is attempting to manipulate care about privacy. However, it also provides a way to limit the sensitivity of the outcome of a mechanism to the choices of any single person, which turns out to be a powerful tool even in the absence of privacy concerns. In this section, we give a brief survey of some of these ideas.

Mechanism Design is the problem of *algorithm design* when the inputs to the algorithm are controlled by individual, self-interested agents, rather than the algorithm designer himself. The algorithm maps its reported inputs to some outcome, over which the agents have preferences. The difficulty is that the agents may mis-report their data if doing so will cause the algorithm to output a different, preferred outcome, and so the mechanism designer must design the algorithm so that the agents are always incentivized to report their true data.

The concerns of mechanism design are very similar to the concerns of private algorithm design. In both cases, the inputs to the algorithm are thought of as belonging to some third party[1] which has preferences over the outcome. In mechanism design, we typically think of individuals as getting some explicit value from the outcomes of the mechanism. In private algorithm design, we typically think of the individual as experiencing some explicit harm from (consequences of) outcomes of the mechanism. Indeed, we can give a utility-theoretic definition of differential privacy which is equivalent to the standard definition, but makes the connection to individual utilities explicit:

Definition 10.1. An algorithm $A : \mathbb{N}^{|\mathcal{X}|} \to R$ is ϵ-differentially private if for every function $f : R \to \mathbb{R}_+$, and for every pair of neighboring databases $x, y \in \mathbb{N}^{|\mathcal{X}|}$:

$$\exp(-\epsilon)\mathbb{E}_{z\sim A(y)}[f(z)] \leq \mathbb{E}_{z\sim A(x)}[f(z)] \leq \exp(\epsilon)\mathbb{E}_{z\sim A(y)}[f(z)].$$

We can think of f as being some function mapping outcomes to an arbitrary agent's utility for those outcomes. With this interpretation, a mechanism is ϵ-differentially private, if for every agent it promises that their participation in the mechanism cannot affect their expected future utility by more than a factor of $\exp(\epsilon)$ *independent of what their utility function might be.*

Let us now give a brief definition of a problem in mechanism design. A mechanism design problem is defined by several objects. There are n agents $i \in [n]$, and a set of outcomes $\mathcal{O}$. Each agent has a type, $t_i \in \mathcal{T}$ which is known only to her, and there is a utility function over outcomes $u : \mathcal{T} \times \mathcal{O} \to [0, 1]$. The utility that agent i gets from an outcome $o \in \mathcal{O}$ is $u(t_i, o)$, which we will often abbreviate as $u_i(o)$. We will write $t \in \mathcal{T}^n$ to denote vectors of all n agent types, with t_i denoting the type of agent i, and $t_{-i} \equiv (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$ denoting the vector of types of all agents *except* agent i. The type of an agent i completely specifies her utility over outcomes — that is, two agents $i \neq j$ such that $t_i = t_j$ will evaluate each outcome identically: $u_i(o) = u_j(o)$ for all $o \in \mathcal{O}$.

[1] In the privacy setting, the database administrator (such as a hospital) might already have access to the data itself, but is nevertheless acting so as to protect the interests of the agents who own the data when it endeavors to protect privacy.

A mechanism M takes as input a set of reported types, one from each player, and selects an outcome. That is, a mechanism is a mapping $M : \mathcal{T}^n \to \mathcal{O}$. Agents will choose to report their types strategically so as to optimize their utility, possibly taking into account what (they think) the other agents will be doing. In particular, they need not report their true types to the mechanism. If an agent is always incentivized to report some type, no matter what her opponents are reporting, then reporting that type is called a *dominant strategy*. If reporting one's true type is a dominant strategy for every agent, then the mechanism is called *truthful*, or equivalently, *dominant strategy truthful*.

Definition 10.2. Given a mechanism $M : \mathcal{T}^n \to \mathcal{O}$, truthful reporting is an ϵ-approximate *dominant strategy* for player i if for every pair of types $t_i, t_i' \in T$, and for every vector of types t_{-i}:

$$u(t_i, M(t_i, t_{-i})) \geq u(t_i, M(t_i', t_{-i})) - \epsilon.$$

If truthful reporting is an ϵ-approximate dominant strategy for every player, we say that M is ϵ-approximately dominant strategy truthful. If $\epsilon = 0$, then M is *exactly truthful*.

That is, a mechanism is truthful if no agent can improve her utility by misrepresenting her type, no matter what the other players report.

Here we can immediately observe a syntactic connection to the definition of differential privacy. We may identify the type space T with the data universe X. The input to the mechanism therefore consists of a database of size n, consisting of the reports of each agent. In fact, when an agent is considering whether she should truthfully report her type t_i or lie, and misreport her type as t_i', she is deciding which of two databases the mechanism should receive: $(t_1, \ldots, t_n)$, or $(t_1, \ldots, t_{i-1}, t_i', t_{i+1}, \ldots, t_n)$. Note that these two databases differ only in the report of agent i! That is, they are *neighboring databases*. Thus, differential privacy gives a guarantee of approximate truthfulness!

10.1 Differential privacy as a solution concept

One of the starting points for investigating the connection between differential privacy and game theory is observing that differential privacy

is a *stronger* condition than approximate truthfulness. Note that for $\epsilon \leq 1$, $\exp(\epsilon) \leq 1 + 2\epsilon$ and so the following proposition is immediate.

Proposition 10.1. If a mechanism M is ϵ-differentially private, then M is also 2ϵ-approximately dominant strategy truthful.

As a solution concept, this has several robustness properties that strategy proof mechanisms do not. By the composition property of differential privacy, the composition of 2 ϵ-differentially private mechanisms remains 4ϵ-approximately dominant strategy truthful. In contrast, the incentive properties of general strategy proof mechanisms may not be preserved under composition.

Another useful property of differential privacy as a solution concept is that it generalizes to group privacy: suppose that t and $t' \in \mathcal{T}^n$ are not neighbors, but instead differ in k indices. Recall that by group privacy we then have for any player i: $\mathbb{E}_{o\sim M(t)}[u_i(o)] \leq \exp(k\epsilon)\mathbb{E}_{o\sim M(t')}[u_i(o)]$. That is, changes in up to k types changes the expected output by at most $\approx (1+k\epsilon)$, when $k \ll 1/\epsilon$. Therefore, differentially private mechanisms make truthful reporting a $2k\epsilon$-approximate dominant strategy *even for coalitions of k agents* — i.e., differential privacy automatically provides robustness to collusion. Again, this is in contrast to general dominant-strategy truthful mechanisms, which in general offer no guarantees against collusion.

Notably, differential privacy allows for these properties in very general settings *without the use of money!* In contrast, the set of exactly dominant strategy truthful mechanisms when monetary transfers are not allowed is extremely limited.

We conclude with a drawback of using differential privacy as a solution concept as stated: not only is truthfully reporting one's type an approximate dominant strategy, *any report* is an approximate dominant strategy! That is, differential privacy makes the outcome approximately independent of any single agent's report. In some settings, this shortcoming can be alleviated. For example, suppose that M is a differentially private mechanism, but that agent utility functions are defined to be functions both of the outcome of the mechanism, *and* of the reported type t_i' of the agent: formally, we view the outcome space as $\mathcal{O}' = \mathcal{O} \times T$. When the agent reports type t_i' to the mechanism, and

the mechanism selects outcome $o \in \mathcal{O}$, then the utility experienced by the agent is controlled by the outcome $o' = (o, t_i')$. Now consider the underlying utility function $u : T \times \mathcal{O}' \to [0, 1]$. Suppose we have that *fixing* a selection o of the mechanism, truthful reporting is a dominant strategy — that is, for all types t_i, t_i', and for all outcomes $o \in \mathcal{O}$:

$$u(t_i, (o, t_i)) \geq u(t_i, (o, t_i')).$$

Then it remains the fact that truthful reporting to an ϵ-differentially private mechanism $M : T^n \to \mathcal{O}$ remains a 2ϵ approximate dominant strategy, because for any misreport t_i' that player i might consider, we have:

$$\begin{aligned} u(t_i, (M(t), t_i)) &= \mathbb{E}_{o \sim M(t)}[u(t_i, (o, t_i))] \\ &\geq (1 + 2\epsilon)\mathbb{E}_{o \sim M(t_i', t_{-i})}[u(t_i, (o, t_i))] \\ &\geq \mathbb{E}_{o \sim M(t_i', t_{-i})}[u(t_i, (o, t_i'))] \\ &= u(t_i, (M(t_i', t_{-i}), t_i')). \end{aligned}$$

However, we no longer have that every report is an approximate dominant strategy, because player i's utility can depend arbitrarily on $o' = (o, t_i')$, and only o (and not player i's report t_i' itself) is differentially private. This will be the case in all examples we consider here.

10.2 Differential privacy as a tool in mechanism design

In this section, we show how the machinery of differential privacy can be used as a tool in designing novel mechanisms.

10.2.1 Warmup: digital goods auctions

To warm up, let us consider a simple special case of the first application of differential privacy in mechanism design. Consider a *digital goods auction*, i.e., one where the seller has an unlimited supply of a good with zero marginal cost to produce, for example a piece of software or other digital media. There are n unit demand buyers for this good, each with unknown valuation $v_i \in [0, 1]$. Informally, the valuation v_i of a bidder i represents the maximum amount of money that buyer i

would be willing to pay for a good. There is no prior distribution on the bidder valuations, so a natural revenue benchmark is the revenue of the *best fixed price*. At a price $p \in [0,1]$, each bidder i with $v_i \geq p$ will buy. Therefore the total revenue of the auctioneer is

$$\mathrm{Rev}(p, v) = p \cdot |\{i : v_i \geq p\}|.$$

The optimal revenue is the revenue of the best fixed price: $\mathrm{OPT} = \max_p \mathrm{Rev}(p, v)$. This setting is well studied: the best known result for exactly dominant strategy truthful mechanisms is a mechanism which achieves revenue at least $\mathrm{OPT} - O(\sqrt{n})$.

We show how a simple application of the exponential mechanism achieves revenue at least $\mathrm{OPT} - O\left(\frac{\log n}{\epsilon}\right)$. That is, the mechanism trades exact for approximate truthfulness, but achieves an exponentially better revenue guarantee. Of course, it also inherits the benefits of differential privacy discussed previously, such as resilience to collusion, and composability.

The idea is to select a price from the exponential mechanism, using as our "quality score" the revenue that this price would obtain. Suppose we choose the range of the exponential mechanism to be $\mathcal{R} = \{\alpha, 2\alpha, \ldots, 1\}$. The size of the range is $|\mathcal{R}| = 1/\alpha$. What have we lost in potential revenue if we restrict ourselves to selecting a price from $\mathcal{R}$? It is not hard to see that

$$\mathrm{OPT}_{\mathcal{R}} \equiv \max_{p \in \mathcal{R}} \mathrm{Rev}(p, v) \geq \mathrm{OPT} - \alpha n.$$

This is because if p^* is the price that achieves the optimal revenue, and we use a price p such that $p^* - \alpha \leq p \leq p^*$, every buyer who bought at the optimal price continues to buy, and provides us with at most α less revenue per buyer. Since there are at most n buyers, the total lost revenue is at most αn.

So how do we parameterize the exponential mechanism? We have a family of discrete ranges $\mathcal{R}$, parameterized by α. For a vector of values v and a price $p \in \mathcal{R}$, we define our quality function to be $q(v, p) = \mathrm{Rev}(v, p)$. Observe that because each value $v_i \in [0,1]$, we can restrict attention to prices $p \leq 1$ and hence, the *sensitivity* of q is $\Delta = 1$: changing one bidder valuation can only change the revenue at a fixed

price by at most $v_i \leq 1$. Therefore, if we require ϵ-differential privacy, by Theorem 3.11, we get that with high probability, the exponential mechanism returns some price p such that

$$\text{Rev}(p, v) \geq (\text{OPT} - \alpha n) - O\left(\frac{1}{\epsilon} \ln\left(\frac{1}{\alpha}\right)\right).$$

Choosing our discretization parameter α to minimize the two sources of error, we find that this mechanism with high probability finds us a price that achieves revenue

$$\text{Rev}(p, v) \geq \text{OPT} - O\left(\frac{\log n}{\epsilon}\right).$$

What is the right level to choose for the privacy parameter ϵ? Note that here, we do not necessarily view privacy itself as a goal of our computation. Rather, ϵ is a way of trading off the revenue guarantee with an upper bound on agent's incentives to deviate. In the literature on large markets in economics, a common goal when exact truthfulness is out of reach is "asymptotic truthfulness" – that is, the maximum incentive that any agent has to deviate from his truthful report tend to 0 as the size of the market n grows large. To achieve a result like that here, all we need to do is set ϵ to be some diminishing function in the number of agents n. For example, if we take $\epsilon = 1/\log(n)$, then we obtain a mechanism that is asymptotically exactly truthful (i.e., as the market grows large, the approximation to truthfulness becomes exact). We can also ask what our approximation to the optimal revenue is as n grows large. Note that our approximation to the optimal revenue is only additive, and so even with this setting of ϵ, we can still guarantee revenue at least $(1 - o(1))\text{OPT}$, so long as OPT grows more quickly than $\log(n)^2$ with the size of the population n.

Finally, notice that we could make the reported value v_i of each agent i binding. In other words, we could allocate an item to agent i and extract payment of the selected posted price p whenever $v_i \geq p$. If we do this, the mechanism is approximately truthful, because the price is picked using a differentially private mechanism. Additionally, it is not the case that *every* report is an approximate dominant strategy: if an agent over-reports, she may be forced to buy the good at a price higher than her true value.

10.2.2 Approximately truthful equilibrium selection mechanisms

We now consider the problem of approximately truthful equilibrium selection. We recall the definition of a *Nash Equilibrium*: Suppose each player has a set of actions $\mathcal{A}$, and can choose to play any action $a_i \in \mathcal{A}$. Suppose, moreover, that *outcomes* are merely choices of actions that the agents might choose to play, and so agent utility functions are defined as $u : \mathcal{T} \times \mathcal{A}^n \to [0, 1]$. Then:

Definition 10.3. A set of actions $a \in \mathcal{A}^n$ is an ϵ-approximate Nash equilibrium if for all players i and for all actions a'_i:

$$u_i(a) \geq u_i(a'_i, a_{-i}) - \epsilon$$

In other words, every agent is simultaneously playing an (approximate) best response to what the other agents are doing, assuming they are playing according to a.

Roughly speaking, the problem is as follows: suppose we are given a game in which each player knows their own payoffs, but not others' payoffs (i.e., the players do not know what the types are of the other agents). The players therefore do not know the equilibrium structure of this game. Even if they did, there might be multiple equilibria, with different agents preferring different equilibria. Can a mechanism offered by an intermediary incentivize agents to truthfully report their utilities and follow the equilibrium it selects?

For example, imagine a city in which (say) Google Navigation is the dominant service. Every morning, each person enters their starting point and destination, receives a set of directions, and chooses his/her route according to those directions. Is it possible to design a navigation service such that: Each agent is incentivized to both (1) report truthfully, and (2) then follow the driving directions provided? Both misreporting start and end points, and truthfully reporting start and end points, but then following a different (shorter) path are to be disincentivized.

Intuitively, our two desiderata are in conflict. In the commuting example above, if we are to guarantee that every player is incentivized to truthfully follow their suggested route, then we must compute an

equilibrium of the game in question given players' reports. On the other hand, to do so, our suggested route to some player i must depend on the reported location/destination pairs of other players. This tension will pose a problem in terms of incentives: if we compute an equilibrium of the game given the reports of the players, an agent can potentially benefit by misreporting, causing us to compute an equilibrium of the wrong game.

This problem would be largely alleviated, however, if the report of agent i only has a tiny effect on the actions of agents $j \neq i$. In this case, agent i could hardly gain an advantage through his effect on other players. Then, assuming that everyone truthfully reported their type, the mechanism would compute an equilibrium of the correct game, and by definition, each agent i could do no better than follow the suggested equilibrium action. In other words, if we could compute an approximate equilibrium of the game under the constraint of *differential privacy*, then truthful reporting, followed by taking the suggested action of the coordination device would be a Nash equilibrium. A moment's reflection reveals that the goal of privately computing an equilibrium is not possible in small games, in which an agent's utility is a highly sensitive function of the actions (and hence utility functions) of the other agents. But what about in large games?

Formally, suppose we have an n player game with action set $\mathcal{A}$, and each agent with type t_i has a utility function $u_i : \mathcal{A}^n \to [0, 1]$. We say that this game is Δ-large if for all players $i \neq j$, vectors of actions $a \in \mathcal{A}^n$, and pairs of actions $a_j, a'_j \in \mathcal{A}$:

$$\left| u_i(a_j, a_{-j}) - u_i(a'_j, a_{-j}) \right| \leq \Delta.$$

In other words, if some agent j unilaterally changes his action, then his affect on the payoff of any other agent $i \neq j$ is at most Δ. Note that if agent j changes his own action, then his payoff can change arbitrarily. Many games are "large" in this sense. In the commuting example above, if Alice changes her route to work she may substantially increase or decrease her commute time, but will only have a minimal impact on the commute time of any other agent Bob. The results in this section are strongest for $\Delta = O(1/n)$, but hold more generally.

First we might ask whether we need privacy at all— could it be the case that in a large game, any algorithm which computes an equilibrium of a game defined by reported types has the stability property that we want? The answer is no. As a simple example, consider n people who must each choose whether to go to the beach (B) or the mountains (M). People privately know their types— each person's utility depends on his own type, his action, and the fraction of other people p who go to the beach. A Beach type gets a payoff of $10p$ if he visits the beach, and $5(1-p)$ if he visits the mountain. A mountain type gets a payoff $5p$ from visiting the beach, and $10(1-p)$ from visiting the mountain. Note that this is a large (i.e., low sensitivity) game — each player's payoffs are insensitive in the actions of others. Further, note that "everyone visits beach" and "everyone visits mountain" are both equilibria of the game, regardless of the realization of types. Consider the mechanism that attempts to implement the following social choice rule— "if the number of beach types is less than half the population, send everyone to the beach, and vice versa." It should be clear that if mountain types are just in the majority, then each mountain type has an incentive to misreport as a beach type; and vice versa. As a result, even though the game is "large" and agents' actions do not affect others' payoffs significantly, simply computing equilibria from reported type profiles does not in general lead to even approximately truthful mechanisms.

Nevertheless, it turns out to be possible to give a mechanism with the following property: it elicits the type t_i of each agent, and then computes an α-approximate correlated equilibrium of the game defined by the reported types.[2] (In some cases, it is possible to strengthen this result to compute an approximate *Nash equilibrium* of the underlying game.) It draws an action profile $a \in \mathcal{A}^n$ from the correlated equilibrium, and reports action a_i to each agent i. The algorithm has the guarantee that simultaneously for all players i, the joint distribution a_{-i} on reports to all players *other than* i is differentially private in

[2] A correlated equilibrium is defined by a joint distribution on profiles of actions, $\mathcal{A}^n$. For an action profile a drawn from the distribution, if agent i is told only a_i, then playing action a_i is a best response given the induced conditional distribution over a_{-i}. An α-approximate correlated equilibrium is one where deviating improves an agent's utility by at most α.

the reported type of agent i. When the algorithm computes a correlated equilibrium of the underlying game, this guarantee is sufficient for a restricted form of approximate truthfulness: agents who have the option to opt-in or opt-out of the mechanism (but not to misreport their type if they opt-in) have no disincentive to opt-out, because no agent i can substantially change the distribution on actions induced on *the other players* by opting out. Moreover, given that he opts in, no agent has incentive not to follow his suggested action, as his suggestion is part of a correlated equilibrium. When the mechanism computes a Nash equilibrium of the underlying game, then the mechanism becomes truthful even when agents have the ability to mis-report their type to the mechanism when they opt in.

More specifically, when these mechanisms compute an α-approximate Nash equilibrium while satisfying ϵ-differential privacy, every agent following the honest behavior (i.e., first opting in and reporting their true type, then following their suggested action) forms an $(2\epsilon + \alpha)$-approximate Nash equilibrium. This is because, by privacy, reporting your true type is a 2ϵ-approximate dominant strategy, and given that everybody reports their true type, the mechanism computes an α-approximate equilibrium of the true game, and hence by definition, following the suggested action is an α-approximate best response. There exist mechanisms for computing and α-approximate equilibrium in large games with $\alpha = O\left(\frac{1}{\sqrt{n\epsilon}}\right)$. Therefore, by setting $\epsilon = O\left(\frac{1}{n^{1/4}}\right)$, this gives an η-approximately truthful equilibrium selection mechanism for

$$\eta = 2\epsilon + \alpha = O\left(\frac{1}{n^{1/4}}\right).$$

In other words, it gives a mechanism for coordinating equilibrium behavior in large games that is asymptotically truthful in the size of the game, all without the need for monetary transfers.

10.2.3 Obtaining exact truthfulness

So far we have discussed mechanisms that are *asymptotically truthful* in large population games. However, what if we want to insist on mechanisms that are *exactly* dominant strategy truthful, while maintaining

some of the nice properties enjoyed by our mechanisms so far: for example, that the mechanisms do not need to be able to extract monetary payments? Can differential privacy help here? It can—in this section, we discuss a framework which uses differentially private mechanisms as a building block toward designing exactly truthful mechanisms without money.

The basic idea is simple and elegant. As we have seen, the exponential mechanism can often give excellent utility guarantees while preserving differential privacy. This doesn't yield an exactly truthful mechanism, but it gives every agent very little incentive to deviate from truthful behavior. What if we could pair this with a second mechanism which need not have good utility guarantees, but gives each agent a strict positive incentive to report truthfully, i.e., a mechanism that essentially only punishes non-truthful behavior? Then, we could randomize between running the two mechanisms. If we put enough weight on the punishing mechanism, then we inherit its strict-truthfulness properties. The remaining weight that is put on the exponential mechanism contributes to the utility properties of the final mechanism. The hope is that since the exponential mechanism is approximately strategy proof to begin with, the randomized mechanism can put small weight on the strictly truthful punishing mechanism, and therefore will have good utility properties.

To design punishing mechanisms, we will have to work in a slightly non-standard environment. Rather than simply picking an outcome, we can model a mechanism as picking an outcome, and then an agent as choosing a *reaction* to that outcome, which together define his utility. Mechanisms will then have the power to *restrict the reactions allowed by the agent based on his reported type*. Formally, we will work in the following framework:

Definition 10.4 (The Environment)**.** An environment is a set N of n players, a set of types $t_i \in \mathcal{T}$, a finite set $\mathcal{O}$ of outcomes, a set of reactions R and a utility function $u : T \times \mathcal{O} \times R \to [0, 1]$.

We write $r_i(t, s, \hat{R}_i) \in \arg\max_{r \in \hat{R}_i} u_i(t, s, r)$ to denote is optimal reaction among choices $\hat{R}_i \subseteq R$ to alternative s if he is of type t.

A direct revelation mechanism $\mathcal{M}$ defines a game which is played as follows:

1. Each player i reports a type $t'_i \in \mathcal{T}$.
2. The mechanism chooses an alternative $s \in \mathcal{O}$ and a subset $\hat{R}_i \subseteq R$ of reactions, for each player i.
3. Each player i chooses a reaction $r_i \in \hat{R}_i$ and experiences utility $u(t_i, s, r_i)$.

Agents play so as to maximize their own utility. Note that since there is no further interaction after the 3rd step, rational agents will pick $r_i = r_i(t_i, s, \hat{R}_i)$, and so we can ignore this as a strategic step. Let $\mathcal{R} = 2^R$. Then a mechanism is a randomized mapping $\mathcal{M} : \mathcal{T} \to \mathcal{O} \times \mathcal{R}^n$.

Let us consider the utilitarian welfare criterion: $F(t, s, r) = \frac{1}{n}\sum_{i=1}^{n} u(t_i, s, r_i)$, Note that this has sensitivity $\Delta = 1/n$, since each agent's utility lies in the range $[0, 1]$. Hence, if we simply choose an outcome s and allow each agent to play their best response reaction, the exponential mechanism is an ϵ-differentially private mechanism, which by Theorem 3.11, achieves social welfare at least $\text{OPT} - O\left(\frac{\log|\mathcal{O}|}{\epsilon n}\right)$ with high probability. Let us denote this instantiation of the exponential mechanism, with quality score F, range $\mathcal{O}$ and privacy parameter ϵ, as $\mathcal{M}_\epsilon$.

The idea is to randomize between the exponential mechanism (with good social welfare properties) and a strictly truthful mechanism which punishes false reporting (but with poor social welfare properties). If we mix appropriately, then we will get an exactly truthful mechanism with reasonable social welfare guarantees.

Here is one such punishing mechanism which is simple, but not necessarily the best for a given problem:

Definition 10.5. The commitment mechanism $M^P(t')$ selects $s \in \mathcal{O}$ uniformly at random and sets $\hat{R}_i = \{r_i(t'_i, s, R_i)\}$, i.e., it picks a random outcome and forces everyone to react as if their reported type was their true type.

Define the *gap* of an environment as

$$\gamma = \min_{i, t_i \neq t'_i, t_{-i}} \max_{s \in \mathcal{O}} \left(u(t_i, s, r_i(t_i, s, R_i)) - u(t_i, s, r_i(t'_i, s, R_i)) \right),$$

i.e., γ is a lower bound over players and types of the worst-case cost (over s) of mis-reporting. Note that for each player, this worst-case is realized with probability at least $1/|\mathcal{O}|$. Therefore we have the following simple observation:

Lemma 10.2. For all i, t_i, t_i', t_{-i}:

$$u(t_i, \mathcal{M}^P(t_i, t_{-i})) \geq u(t_i, \mathcal{M}^P(t_i', t_{-i})) + \frac{\gamma}{|\mathcal{O}|}.$$

Note that the commitment mechanism is strictly truthful: every individual has at least a $\frac{\gamma}{|\mathcal{O}|}$ incentive not to lie.

This suggests an exactly truthful mechanism with good social welfare guarantees:

Definition 10.6. The punishing exponential mechanism $\mathcal{M}_\epsilon^P(t)$ defined with parameter $0 \leq q \leq 1$ selects the exponential mechanism $\mathcal{M}_\epsilon(t)$ with probability $1-q$ and the punishing mechanism $\mathcal{M}^P(t)$ with complementary probability q.

Observe that by linearity of expectation, we have for all t_i, t_i', t_{-i}:

$$\begin{aligned}
u(t_i, \mathcal{M}_\epsilon^P(t_i, t_{-i})) &= (1-q) \cdot u(t_i, \mathcal{M}_\epsilon(t_i, t_{-i})) + q \cdot u(t_i, \mathcal{M}^P(t_i, t_{-i})) \\
&\geq (1-q)\left(u(t_i, \mathcal{M}_\epsilon(t_i', t_{-i})) - 2\epsilon\right) \\
&\quad + q\left(u(t_i, \mathcal{M}^P(t_i', t_{-i})) + \frac{\gamma}{|\mathcal{O}|}\right) \\
&= u(t_i, \mathcal{M}_\epsilon^P(t_i', t_{-i})) - (1-q)2\epsilon + q\frac{\gamma}{|\mathcal{O}|} \\
&= u(t_i, \mathcal{M}_\epsilon^P(t_i', t_{-i})) - 2\epsilon + q\left(2\epsilon + \frac{\gamma}{|\mathcal{O}|}\right).
\end{aligned}$$

The following two theorems show incentive and social welfare properties of this mechanism.

Theorem 10.3. If $2\epsilon \leq \frac{q\gamma}{|\mathcal{O}|}$ then $\mathcal{M}_\epsilon^P$ is strictly truthful.

Note that we also have utility guarantees for this mechanism. Setting the parameter q so that we have a truthful mechanism:

$$\begin{aligned}
&\mathbb{E}_{s,\hat{R}\sim\mathcal{M}_\epsilon^P}[F(t,s,r(t,s,\hat{R}))] \\
&\quad\geq (1-q)\cdot\mathbb{E}_{s,\hat{R}\sim\mathcal{M}_\epsilon}[F(t,s,r(t,s,\hat{R}))] \\
&\quad= \left(1-\frac{2\epsilon|\mathcal{O}|}{\gamma}\right)\cdot\mathbb{E}_{s,\hat{R}\sim\mathcal{M}_\epsilon}[F(t,s,r(t,s,\hat{R}))] \\
&\quad\geq \left(1-\frac{2\epsilon|\mathcal{O}|}{\gamma}\right)\cdot\left(\max_{t,s,r}F(t,s,r)-O\left(\frac{1}{\epsilon n}\log|\mathcal{O}|\right)\right) \\
&\quad\geq \max_{t,s,r}F(t,s,r)-\frac{2\epsilon|\mathcal{O}|}{\gamma}-O\left(\frac{1}{\epsilon n}\log|\mathcal{O}|\right).
\end{aligned}$$

Setting

$$\epsilon\in O\left(\sqrt{\frac{\log|\mathcal{O}|\gamma}{|\mathcal{O}|n}}\right)$$

we find:

$$\mathbb{E}_{s,\hat{R}\sim\mathcal{M}_\epsilon^P}[F(t,s,r(t,s,\hat{R}))]\geq\max_{t,s,r}F(t,s,r)-O\left(\sqrt{\frac{|\mathcal{O}|\log|\mathcal{O}|}{\gamma n}}\right).$$

Note that in this calculation, we assume that $\epsilon\leq\gamma/(2|\mathcal{O}|)$ so that $q=\frac{2\epsilon|\mathcal{O}|}{\gamma}\leq 1$ and the mechanism is well defined. This is true for sufficiently large n. That is, we have shown:

Theorem 10.4. For sufficiently large n, M_ϵ^P achieves social welfare at least

$$\mathrm{OPT}-O\left(\sqrt{\frac{|\mathcal{O}|\log|\mathcal{O}|}{\gamma n}}\right).$$

Note that this mechanism is truthful without the need for payments!

Let us now consider an application of this framework: the facility location game. Suppose that a city wants to build k hospitals to minimize the average distance between each citizen and their closest hospital. To simplify matters, we make the mild assumption that the city is built on a discretization of the unit line.[3] Formally, let

[3] If this is not the case, we can easily raze and then re-build the city.

$L(m) = \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$ denote the discrete unit line with step-size $1/m$. $|L(m)| = m+1$. Let $\mathcal{T} = R_i = L(m)$ for all i and let $|\mathcal{O}| = L(m)^k$. Define the utility of agent i to be:

$$u(t_i, s, r_i) = \begin{cases} -|t_i - r_i|, & \text{If } r_i \in s; \\ -1, & \text{otherwise.} \end{cases}$$

In other words, agents are associated with points on the line, and an outcome is an assignment of a location on the line to each of the k facilities. Agents can react to a set of facilities by deciding which one to go to, and their cost for such a decision is the distance between their own location (i.e., their type) and the facility that they have chosen. Note that $r_i(t_i, s)$ is here the closest facility $r_i \in s$.

We can instantiate Theorem 10.4. In this case, we have: $|\mathcal{O}| = (m+1)^k$ and $\gamma = 1/m$, because any two positions $t_i \neq t_i'$ differ by at least $1/m$. Hence, we have:

Theorem 10.5. M_ϵ^P instantiated for the facility location game is strictly truthful and achieves social welfare at least:

$$\text{OPT} - O\left(\sqrt{\frac{km(m+1)^k \log m}{n}}\right).$$

This is already very good for small numbers of facilities k, since we expect that $\text{OPT} = \Omega(1)$.

10.3 Mechanism design for privacy aware agents

In the previous section, we saw that differential privacy can be useful as a tool to design mechanisms, *for agents who care only about the outcome chosen by the mechanism*. We here primarily viewed privacy as a tool to accomplish goals in traditional mechanism design. As a side affect, these mechanisms also preserved the privacy of the reported player types. Is this itself a worthy goal? *Why* might we want our mechanisms to preserve the privacy of agent types?

A bit of reflection reveals that agents might care about privacy. Indeed, basic introspection suggests that in the real world, agents value the ability to keep certain "sensitive" information private, for example,

health information or sexual preferences. In this section, we consider the question of how to model this value for privacy, and various approaches taken in the literature.

Given that agents might have preferences for privacy, it is worth considering the design of mechanisms that preserve privacy *as an additional goal*, even for tasks such as welfare maximization that we can already solve non-privately. As we will see, it is indeed possible to generalize the VCG mechanism to *privately* approximately optimize social welfare in *any* social choice problem, with a smooth trade-off between the privacy parameter and the approximation parameter, all while guaranteeing exact dominant strategy truthfulness.

However, we might wish to go further. In the presence of agents with preferences for privacy, if we wish to design truthful mechanisms, we must somehow model their preferences for privacy in their utility function, and then design mechanisms which are truthful with respect to these new "privacy aware" utility functions. As we have seen with differential privacy, it is most natural to model privacy as a property of the mechanism itself. Thus, our utility functions are not merely functions of the outcome, but functions of the outcome and of the mechanism itself. In almost all models, agent utilities for outcomes are treated as linearly separable, that is, we will have for each agent i,

$$u_i(o, \mathcal{M}, t) \equiv \mu_i(o) - c_i(o, \mathcal{M}, t).$$

Here $\mu_i(o)$ represents agent is utility for outcome o and $c_i(o, \mathcal{M}, t)$ the (privacy) cost that agent i experiences when outcome o is chosen with mechanism $\mathcal{M}$.

We will first consider perhaps the simplest (and most naïve) model for the privacy cost function c_i. Recall that for $\epsilon \ll 1$, differential privacy promises that for each agent i, and for every possible utility function f_i, type vector $t \in \mathcal{T}^n$, and deviation $t'_i \in \mathcal{T}$:

$$|\mathbb{E}_{o \sim M(t_i, t_{-i})}[f_i(o)] - \mathbb{E}_{o \sim M(t'_i, t_{-i})}[f_i(o)]| \leq 2\epsilon \mathbb{E}_{o \sim M(t)}[f_i(o)].$$

If we view f_i as representing the "expected future utility" for agent i, it is therefore natural to model agent i's cost for having his data used in an ϵ-differentially private computation as being linear in ϵ. That is,

we think of agent i as being parameterized by some value $v_i \in \mathbb{R}$, and take:

$$c_i(o, \mathcal{M}, t) = \epsilon v_i,$$

where ϵ is the smallest value such that $\mathcal{M}$ is ϵ-differentially private. Here we imagine v_i to represent a quantity like $\mathbb{E}_{o \sim M(t)}[f_i(o)]$. In this setting, c_i does not depend on the outcome o or the type profile t.

Using this naïve privacy measure, we discuss a basic problem in private data analysis: how to collect the data, when the owners of the data value their privacy and insist on being compensated for it. In this setting, there is no "outcome" that agents value, other than payments, there is only dis-utility for privacy loss. We will then discuss shortcomings of this (and other) measures of the dis-utility for privacy loss, as well as privacy in more general mechanism design settings when agents *do* have utility for the outcome of the mechanism.

10.3.1 A private generalization of the VCG mechanism

Suppose we have a general social choice problem, defined by an outcome space $\mathcal{O}$, and a set of agents N with arbitrary preferences over the outcomes given by $u_i : \mathcal{O} \to [0, 1]$. We might want to choose an outcome $o \in \mathcal{O}$ to maximize the *social welfare* $F(o) = \frac{1}{n} \sum_{i=1}^{n} u_i(o)$. It is well known that in any such setting, the *VCG* mechanism can implement the outcome o^* which exactly maximizes the social welfare, while charging payments that make truth-telling a dominant strategy. What if we want to achieve the same result, while also preserving privacy? How must the privacy parameter ϵ trade off with our approximation to the optimal social welfare?

Recall that we could use the exponential mechanism to choose an outcome $o \in \mathcal{O}$, with quality score F. For privacy parameter ϵ, this would give a distribution $\mathcal{M}_\epsilon$ defined to be $\Pr[\mathcal{M}_\epsilon = o] \propto \exp\left(\frac{\epsilon F(o)}{2n}\right)$. Moreover, this mechanism has good social welfare properties: with probability $1 - \beta$, it selects some o such that: $F(o) \geq F(o^*) - \frac{2}{\epsilon n}\left(\ln \frac{|\mathcal{O}|}{\beta}\right)$. But as we saw, differential privacy only gives ϵ-approximate truthfulness.

However, it can be shown that $\mathcal{M}_\epsilon$ is the solution to the following exact optimization problem:

$$\mathcal{M}_\epsilon = \arg\max_{\mathcal{D}\in\Delta\mathcal{O}} \left(\mathbb{E}_{o\sim\mathcal{D}}[F(o)] + \frac{2}{\epsilon n} H(\mathcal{D}) \right),$$

where H represents the *Shannon Entropy* of the distribution $\mathcal{D}$. In other words, the exponential mechanism is the distribution which exactly maximizes the expected social welfare, *plus* the entropy of the distribution weighted by $2/(\epsilon n)$. This is significant for the following reason: it is known that any mechanism that *exactly* maximizes expected player utilities in any finite range (known as maximal in distributional range mechanisms) can be paired with payments to be made exactly dominant strategy truthful. The exponential mechanism is the distribution that *exactly* maximizes expected social welfare, plus entropy. In other words, if we imagine that we have added a single additional player whose utility is exactly the entropy of the distribution, then the exponential mechanism is maximal in distributional range. Hence, it can be paired with payments that make truthful reporting a dominant strategy for all players — in particular, for the n real players. Moreover, it can be shown how to charge payments in such a way as to preserve privacy. The upshot is that for any social choice problem, the social welfare can be approximated in a manner that both preserves differential privacy, and is exactly truthful.

10.3.2 The sensitive surveyor's problem

In this section, we consider the problem of a data analyst who wishes to conduct a study using the private data of a collection of individuals. However, he must *convince* these individuals to hand over their data! Individuals experience costs for privacy loss. The data analyst can mitigate these costs by guaranteeing differential privacy and compensating them for their loss, while trying to get a representative sample of data.

Consider the following stylized problem of the sensitive surveyor Alice. She is tasked with conducting a survey of a set of n individuals N, to determine what proportion of the individuals $i \in N$ satisfy some property $P(i)$. Her ultimate goal is to discover the true value of this statistic, $s = \frac{1}{n}|\{i \in N : P(i)\}|$, but if that is not possible, she will be

satisfied with some estimate $\hat{s}$ such that the error, $|\hat{s} - s|$, is minimized. We will adopt a notion of accuracy based on large deviation bounds, and say that a surveying mechanism is α-accurate if $\Pr[|\hat{s} - s| \geq \alpha] \leq \frac{1}{3}$. The inevitable catch is that individuals value their privacy and will not participate in the survey for free. Individuals experience some *cost* as a function of their loss in privacy when they interact with Alice, and must be compensated for this loss. To make matters worse, these individuals are rational (i.e., selfish) agents, and are apt to misreport their costs to Alice if doing so will result in a financial gain. This places Alice's problem squarely in the domain of mechanism design, and requires Alice to develop a scheme for trading off statistical accuracy with cost, all while managing the incentives of the individuals.

As an aside, this stylized problem is broadly relevant to any organization that makes use of collections of potentially sensitive data. This includes, for example, the use of search logs to provide search query completion and the use of browsing history to improve search engine ranking, the use of social network data to select display ads and to recommend new links, and the myriad other data-driven services now available on the web. In all of these cases, value is being derived from the statistical properties of a collection of sensitive data in exchange for some payment.[4]

Collecting data in exchange for some fixed price could lead to a biased estimate of population statistics, because such a scheme will result in collecting data only from those individuals who value their privacy less than the price being offered. However, without interacting with the agents, we have no way of knowing what price we can offer so that we will have broad enough participation to guarantee that the answer we collect has only small bias. To obtain an accurate estimate of the statistic, it is therefore natural to consider buying private data using an auction — as a means of discovering this price. There are two obvious obstacles which one must confront when conducting an auction for private data, and an additional obstacle which is less obvious but more insidious. The first obstacle is that one must have a quantitative

[4]The payment need not be explicit and/or dollar denominated — for example, it may be the use of a "free" service.

formalization of "privacy" which can be used to measure agents' costs under various operations on their data. Here, differential privacy provides an obvious tool. For small values of ϵ, because $\exp(\epsilon) \approx (1 + \epsilon)$, and so as discussed earlier, a simple (but possibly naive) first cut at a model is to view each agent as having some *linear* cost for participating in a private study. We here imagine that each agent i has an unknown value for privacy v_i, and experiences a cost $c_i(\epsilon) = \epsilon v_i$ when his private data is used in an ϵ-differentially private manner.[5] The second obstacle is that our objective is to trade off with *statistical accuracy*, and the latter is not well-studied objective in mechanism design.

The final, more insidious obstacle, is that an individual's cost for privacy loss may be highly correlated with his private data itself! Suppose we only know Bob has a high value for privacy of his AIDS status, but do not explicitly know his AIDS status itself. This is already disclosive because Bob's AIDS status is likely correlated with his value for privacy, and knowing that he has a high cost for privacy lets us update our belief about what his private data might be. More to the point, suppose that in the first step of a survey of AIDS prevalence, we ask each individual to report their value for privacy, with the intention of then running an auction to choose which individuals to buy data from. If agents report truthfully, we may find that the reported values naturally form two clusters: low value agents, and high value agents. In this case, we may have learned something about the population statistic even before collecting any data or making any payments— and therefore, the agents will have already experienced a cost. As a result, the agents may misreport their value, which could introduce a bias in the survey results. This phenomenon makes direct revelation mechanisms problematic, and distinguishes this problem from classical mechanism design.

Armed with a means of quantifying an agent i's loss for allowing his data to be used by an ϵ-differentially-private algorithm ($c_i(\epsilon) = \epsilon v_i$), we are almost ready to describe results for the sensitive surveyor's problem. Recall that a differentially private algorithm is some mapping $M : \mathcal{T}^n \to \mathcal{O}$, for a general type space $\mathcal{T}$. It remains to define what

[5] As we will discuss later, this assumption can be problematic.

exactly the type space $\mathcal{T}$ is. We will consider two models. In both models, we will associate with each individual a bit $b_i \in \{0,1\}$ which represents whether they satisfy the sensitive predicate $P(i)$, as well as a value for privacy $v_i \in \mathbb{R}^+$.

1. In the *insensitive value model*, we calculate the ϵ parameter of the private mechanism by letting the type space be $\mathcal{T} = \{0,1\}$: i.e., we measure privacy cost only with respect to how the mechanism treats the sensitive bit b_i, and ignore how it treats the reported values for privacy, v_i.[6]
2. In the *sensitive value model*, we calculate the ϵ parameter of the private mechanism by letting the type space be $\mathcal{T} = (\{0,1\} \times \mathbb{R}^+)$: i.e., we measure privacy with respect to how it treats the pair (b_i, v_i) for each individual.

Intuitively, the insensitive value model treats individuals as ignoring the potential privacy loss due to correlations between their values for privacy and their private bits, whereas the sensitive value model treats individuals as assuming these correlations are worst-case, i.e., their values v_i are just as disclosive as their private bits b_i. It is known that in the insensitive value model, one can derive approximately optimal direct revelation mechanisms that achieve high accuracy and low cost. By contrast, in the *sensitive value model*, no individually rational direct revelation mechanism can achieve any non-trivial accuracy.

This leaves a somewhat unsatisfying state of affairs. The sensitive value model captures the delicate issues that we really want to deal with, and yet there we have an impossibility result! Getting around this result in a satisfying way (e.g., by changing the model, or the powers of the mechanism) remains an intriguing open question.

10.3.3 Better measures for the cost of privacy

In the previous section, we took the naive modeling assumption that the cost experienced by participation in an ϵ-differentially private mechanism M was $c_i(o, \mathcal{M}, t) = \epsilon v_i$ for some numeric value v_i. This measure

[6] That is, the part of the mapping dealing with reported values need not be differentially private.

is problematic for several reasons. First, although differential privacy promises that any agent's loss in utility is *upper bounded* by a quantity that is (approximately) linear in ϵ, there is no reason to believe that agents' costs are *lower bounded* by such a quantity. That is, while taking $c_i(o, \mathcal{M}, t) \leq \epsilon v_i$ is well motivated, there is little support for making the inequality an equality. Second, (it turns out) *any* privacy measure which is a deterministic function only of ϵ (not just a linear function) leads to problematic behavioral predictions.

So how else might we model c_i? One natural measure is the *mutual information* between the reported type of agent i, and the outcome of the mechanism. For this to be well defined, we must be in a world where each agent's type t_i is drawn from a known prior, $t_i \sim \mathcal{T}$. Each agent's strategy is a mapping $\sigma_i : \mathcal{T} \to \mathcal{T}$, determining what type he reports, given his true type. We could then define

$$c_i(o, \mathcal{M}, \sigma) = I(\mathcal{T}; \mathcal{M}(t_{-i}, \sigma(\mathcal{T})),$$

where I is the mutual information between the random variable $\mathcal{T}$ representing the prior on agent is type, and $\mathcal{M}(t_{-i}, \sigma(\mathcal{T}))$, the random variable representing the outcome of the mechanism, given agent is strategy.

This measure has significant appeal, because it represents how "related" the output of the mechanism is to the true type of agent i. However, in addition to requiring a prior over agent types, observe an interesting paradox that results from this measure of privacy loss. Consider a world in which there are two kinds of sandwich breads: Rye (R), and Wheat (W). Moreover, in this world, sandwich preferences are highly embarrassing and held private. The prior on types $\mathcal{T}$ is uniform over R and W, and the mechanism $\mathcal{M}$ simply gives agent i a sandwich of the type that he purports to prefer. Now consider two possible strategies, σ_{truthful} and σ_{random}. σ_{truthful} corresponds to truthfully reporting sandwich preferences (and subsequently leads to eating the preferred sandwich type), while σ_{random} randomly reports independent of true type (and results in the preferred sandwich only half the time). The cost of using the random strategy is $I(\mathcal{T}; \mathcal{M}(t_{-i}, \sigma_{\text{random}}(\mathcal{T})) = 0$, since the output is independent of agent i's type. On the other hand, the cost of truthfully reporting is $I(\mathcal{T}; \mathcal{M}(t_{-i}, \sigma_{\text{truthful}}(\mathcal{T})) = 1$, since

the sandwich outcome is now the identity function on agent is type. However, from the perspective of any outside observer, the two strategies are indistinguishable! In both cases, agent i receives a uniformly random sandwich. Why then should anyone choose the random strategy? So long as an adversary *believes* they are choosing randomly, they should choose the honest strategy.

Another approach, which does not need a prior on agent types, is as follows. We may model agents as having a cost function c_i that satisfies:

$$|c_i(o, \mathcal{M}, t)| = \ln \left(\max_{t_i, t_i' \in \mathcal{T}} \frac{\Pr[\mathcal{M}(t_i, t_{-i}) = o]}{\Pr[\mathcal{M}(t_i', t_{-i}) = o]} \right).$$

Note that if $\mathcal{M}$ is ϵ-differentially private, then

$$\max_{t \in \mathcal{T}^n} \max_{o \in \mathcal{O}} \max_{t_i, t_i' \in \mathcal{T}} \ln \left(\frac{\Pr[\mathcal{M}(t_i, t_{-i}) = o]}{\Pr[\mathcal{M}(t_i', t_{-i}) = o]} \right) \leq \epsilon.$$

That is, we can view differential privacy as bounding the *worst-case* privacy loss over all possible outcomes, whereas the measure proposed here considers only the privacy loss for the outcome o (and type vector t) actually realized. Thus, for any differentially private mechanism $\mathcal{M}$, $|c_i(o, \mathcal{M}, t)| \leq \epsilon$ for all o, t, but it will be important that the cost can vary by outcome.

We can then consider the following allocation rule for maximizing social welfare $F(o) = \sum_{i=1}^{n} u_i(o)$.[7] We discuss the case when $|\mathcal{O}| = 2$ (which does not require payments), but it is possible to analyze the general case (with payments), which privately implements the VCG mechanism for any social choice problem.

1. For each outcome $o \in \mathcal{O}$, choose a random number r_o from the distribution $\Pr[r_o = x] \propto \exp(-\epsilon|x|)$.
2. Output $o^* = \arg\max_{o \in \mathcal{O}}(F(o) + r_o)$.

The above mechanism is ϵ-differentially private, and that it is truthful for privacy aware agents, so long as for each agent i, and for the two outcomes $o, o' \in \mathcal{O}$, $|\mu_i(o) - \mu_i(o')| > 2\epsilon$. Note that this will be true

[7] This allocation rule is extremely similar to, and indeed can be modified to be identical to the exponential mechanism.

for small enough ϵ so long as agent utilities for outcomes are distinct. The analysis proceeds by considering an arbitrary fixed realization of the random variables r_o, and an arbitrary deviation t_i' from truthful reporting for the ith agent. There are two cases: In the first case, the deviation does not change the outcome o of the mechanism. In this case, *neither* the agent's utility for the outcome μ_i, nor his cost for privacy loss c_i change at all, and so the agent does not benefit from deviating. In the second case, if the outcome changes from o to o' when agent i deviates, it must be that $\mu_i(o') < \mu_i(o) - 2\epsilon$. By differential privacy, however, $|c_i(o, \mathcal{M}, t) - c_i(o', \mathcal{M}, t)| \leq 2\epsilon$, and so the change in privacy cost cannot be enough to make it beneficial.

Finally, the most conservative approach to modeling costs for privacy generally considered is as follows. Given an ϵ-differentially private mechanism $\mathcal{M}$, assume only that

$$c_i(o, \mathcal{M}, t) \leq \epsilon v_i,$$

for some number v_i. This is similar to the linear cost functions that we considered earlier, but crucially, here we assume only an upper bound. This assumption is satisfied by all of the other models for privacy cost that we have considered thus far. It can be shown that many mechanisms that combine a differentially private algorithm with a punishing mechanism that has the ability to restrict user choices, like those that we considered in Section 10.2.3, maintain their truthfulness properties in the presence of agents with preferences for privacy, so long as the values v_i are bounded.

10.4 Bibliographical notes

This section is based off of a survey of Pai and Roth [70] and a survey of Roth [73]. The connections between differential privacy and mechanism design were first suggested by Jason Hartline and investigated by McSherry and Talwar in their seminal work, "Mechanism Design via Differential Privacy" [61], where they considered the application of differential privacy to designing approximately truthful digital goods auctions. The best result for exactly truthful mechanisms in the digital goods setting is due to Balcan et al. [2].

The problem of designing exactly truthful mechanisms using differential privacy as a tool was first explored by Nissim, Smorodinsky, and Tennenholtz in [69], who also first posed a criticism as differential privacy (by itself) used as a solution concept. The example in this section of using differential privacy to obtain exactly truthful mechanisms is taken directly from [69]. The sensitive surveyors problem was first considered by Ghosh and Roth [36], and expanded on by [56, 34, 75, 16]. Fleischer and Lyu [34] consider the Bayesian setting discussed in this section, and Ligett and Roth [56] consider the worst-case setting with take-it-or-leave-it offers, both in an attempt to get around the impossibility result of [36]. Ghosh and Ligett consider a related model in which participation decisions (and privacy guarantees) are determined only in equilibrium [35].

The question of conducting mechanism design in the presence of agents who explicitly value privacy as part of their utility function was first raised by the influential work of Xiao [85], who considered (among other measures for privacy cost) the mutual information cost function. Following this, Chen et al. [15] and Nissim et al. [67] showed how in two distinct models, truthful mechanisms can sometimes be designed even for agents who value privacy. Chen Chong, Kash, Moran, and Vadhan considered the outcome-based cost function that we discussed in this section, and Nissim, Orlandi, and Smorodinsky considered the conservative model of only upper bounding each agent's cost by a linear function in $\epsilon>$ The "sandwich paradox" of valuing privacy according to mutual information is due to Nissim, Orlandi, and Smorodinsky.

Huang and Kannan proved that the exponential mechanism could be made exactly truthful with the addition of payments [49]. Kearns Pai, Roth, and Ullman showed how differential privacy could be used to derive asymptotically truthful equilibrium selection mechanisms [54] by privately computing correlated equilibria in large games. These results were strengthened by Rogers and Roth [71], who showed how to privately compute approximate *Nash* equilibria in large congestion games, which leads to stronger incentive properties of the mechanism. Both of these papers use the solution concept of "Joint Differential Privacy,"

which requires that for every player i, the joint distribution on messages sent to *other* players $j \neq i$ be differentially private in is report. This solution concept has also proven useful in other settings of private mechanism design settings, including an algorithm for computing private matchings by Hsu et al. [47].