Bayesian Mechanism Design with Efficiency, Privacy, and Approximate Truthfulness*

Samantha Leung and Edward Lui

Department of Computer Science, Cornell University
{samlyy,luied}@cs.cornell.edu

Abstract. Recently, there has been a number of papers relating mechanism design and privacy (e.g., see [1–6]). All of these papers consider a worst-case setting where there is no probabilistic information about the players' types. In this paper, we investigate mechanism design and privacy in the *Bayesian* setting, where the players' types are drawn from some common distribution. We adapt the notion of *differential privacy* to the Bayesian mechanism design setting, obtaining *Bayesian differential privacy*. We also define a robust notion of approximate truthfulness for Bayesian mechanisms, which we call *persistent approximate truthfulness*. We give several classes of mechanisms (e.g., social welfare mechanisms and histogram mechanisms) that achieve both Bayesian differential privacy and persistent approximate truthfulness. These classes of mechanisms can achieve optimal (economic) efficiency, and do not use any payments. We also demonstrate that by considering the above mechanisms in a modified mechanism design model, the above mechanisms can achieve actual truthfulness.

1 Introduction

One of the main goals in mechanism design is to design mechanisms that achieve a socially desirable outcome even if the players behave selfishly. Because of the revelation principle, mechanism design has focused on direct (revelation) mechanisms where each player simply reports his/her private type (or valuation). This leads to the issue of privacy, where the players may be concerned that the mechanism's output may leak information about their private types (even if the mechanism is trusted).

Mechanism Design and Privacy. Traditional mechanism design did not include the aspect of privacy. However, in the context of releasing information from databases, the issue of privacy has already been studied quite extensively. In this context, the current standard notion of privacy is *differential privacy* [7, 8]. A data release algorithm satisfies differential privacy if the algorithm's output distribution does not change much when one person's data is changed in the

* Work supported in part by NSF grants IIS-0534064, IIS-0812045, IIS-0911036, CCF-0746990; AFOSR grants FA9550-08-1-0438, FA9550-09-1-0266, FA9550-10-1-0093; ARO grant W911NF-09-1-0281. We thank Joseph Y. Halpern for helpful discussions.

database. This implies that the algorithm does not leak much information about any person in the database.

Recently, there has been a number of papers that combine mechanism design with differential privacy. In [1], McSherry and Talwar develop a general mechanism called the *exponential mechanism* that is differentially private; they also show that any differentially private mechanism is *approximately* truthful. In [4], Nissim, Smorodinsky, and Tennenholtz modify the standard mechanism design model by adding a "reaction stage"; in this new model, the authors combine differentially private mechanisms with a "punishing mechanism" to obtain mechanisms that are *actually* truthful. However, the mechanisms in [4] might not protect the privacy of the players, due to the reaction stage.

The main goal of the above two papers was to use differential privacy as a tool for achieving some form of truthfulness, as opposed to achieving privacy for the players. However, there has been other papers that focus on designing mechanisms that protect the privacy of the players. In [6], Huang and Kannan show that a pricing scheme can be added to the exponential mechanism to make it *actually* truthful, resulting in a general mechanism that is both differentially private and truthful. In [2], Xiao provides a transformation that takes truthful mechanisms and transforms them into truthful and differentially private mechanisms. On the other hand, Xiao also shows that a mechanism that is truthful and differentially private might not be truthful in a model where the players are "privacy-aware", i.e., privacy is explicitly captured in the players' utility functions. In [3], Chen et al. construct mechanisms that are truthful even in a model where the players are privacy-aware. In [5], Nissim, Orlandi, and Smorodinsky construct mechanisms that are truthful in a different privacy-aware model.

Bayesian Mechanism Design. One desirable property of a mechanism is (economic) *efficiency*; in fact, it would be best if the mechanism always chooses a social alternative that is *optimal* with respect to some measure of efficiency, such as social welfare. However, such *optimal efficiency* is not achieved by any of the above results. In fact, it is not possible for a differentially private mechanism to achieve optimal efficiency (for a non-trivial problem), since the mechanism has to be randomized in order to satisfy differential privacy. However, all of the above results are in a worst-case setting where there is no probabilistic information about the players' types. If we consider a non-worst-case setting, then it may be possible for a mechanism to achieve differential privacy without using any randomization.

One such setting is the *Bayesian* setting, where the players' types are drawn from some common distribution. Such a setting follows the Bayesian approach that has been the standard in economic theory for many decades. Recently, mechanism design in the Bayesian setting has also been gaining popularity in the computer science community. Thus, it is interesting to consider the issue of privacy in the Bayesian setting as well. In particular, it may be possible for a Bayesian mechanism to achieve optimal efficiency while satisfying some form of differential privacy. Achieving optimal efficiency may be critical for certain problems, such as presidential elections and kidney transplant allocations, where it

may be unethical and/or unfair to make a non-optimal choice. Although differentially private mechanisms in the worst-case setting may asymptotically achieve nearly optimal efficiency in expectation (or with reasonably high probability), there is no guarantee that the chosen outcome for a particular execution of the mechanism is actually close to optimal.

Bayesian Differential Privacy and Persistent Approximate Truthfulness. In this paper, we consider mechanism design in the Bayesian setting, and our main goal is to construct useful mechanisms that achieve optimal efficiency, some form of differential privacy, and some notion of truthfulness. Since differential privacy is a worst-case notion in the sense that no distributional assumptions are made on the input of the mechanism, we first adapt the notion of differential privacy to the Bayesian mechanism design setting. We call this new notion *Bayesian differential privacy*; this is the privacy notion that we use in this paper.

As mentioned above, Xiao [2] showed that a mechanism that is truthful and differentially private might not be truthful in a model where privacy is explicitly captured in the players' utility functions. In this paper, we do not use such a model, since there are many settings where the players would already be satisfied with differential privacy and would not report strategically in an attempt to further protect their privacy. Our results will be meaningful in these settings; furthermore, even in a setting where we want to explicitly capture privacy in the players' utility functions, our techniques and results can still be useful in constructing truthful mechanisms (similar to how the mechanisms in [3] and [5] are still based on differentially private mechanisms).

We also want our mechanisms to satisfy some form of truthfulness. The standard notion of truthfulness in Bayesian mechanism design is that the truthful strategy profile is a Bayes-Nash equilibrium. Similar to [1], we first relax truthfulness so that the truthful strategy profile only needs to be an *ϵ-Bayes-Nash equilibrium*, where an ϵ margin is allowed in the Nash conditions. However, we would like to obtain notions of truthfulness that are stronger than that provided by the ϵ-Bayes-Nash equilibrium. Thus, we strengthen the ϵ-Bayes-Nash equilibrium such that even if up to k players deviate from the equilibrium, everyone else's best-response is still to adhere to their part of the equilibrium. We call this new equilibrium concept the *k-tolerant ϵ-Bayes-Nash equilibrium*. We would also like our equilibrium concept to be resilient against coalitions. Thus, we further strengthen our notion of k-tolerant ϵ-Bayes-Nash equilibrium to *(k, r)-persistent ϵ-Bayes-Nash equilibrium*, which is resilient against coalitions of size r even in the presence of k deviating players. The notion of truthfulness we use requires that the truthful strategy profile is a (k, r)-persistent ϵ-Bayes-Nash equilibrium, which we will refer to as *persistent approximate truthfulness*.

1.1 Our Results

In this paper, we present three classes of mechanisms that achieve both Bayesian differential privacy and persistent approximate truthfulness:

Histogram Mechanisms. Roughly speaking, a histogram mechanism is a mechanism that first computes a histogram from the reported types, and then chooses a social alternative based only on the histogram. In Section 4.1, we show that if every bin of the histogram has positive expected count, then the histogram mechanism is both Bayesian differentially private and persistent approximately truthful.

Mechanisms for Two Social Alternatives. Roughly speaking, this class includes any mechanism that makes a choice between two social alternatives $\{A, B\}$ based on the difference between the sums of two functions $u(\cdot, A)$ and $u(\cdot, B)$ on the types. In Section 4.2, we show that as long as the random variable $u(t, A) - u(t, B)$ (where t is distributed according to the type distribution) has non-zero variance, then such a mechanism is both Bayesian differentially private and persistent approximately truthful.

Social Welfare Mechanisms. Roughly speaking, this class includes any mechanism that makes a choice based on the social welfare provided by each social alternative. An important subset of these mechanisms is the set of mechanisms that maximize social welfare. In Section 4.3, we show that if the players' valuations for each social alternative are normally distributed, then such a mechanism is both Bayesian differentially private and persistent approximately truthful. In our full paper, we generalize this result to the case where the players' valuations for each social alternative are arbitrarily distributed with non-zero variance.

The mechanisms in the above three classes are all deterministic and can achieve optimal efficiency. Furthermore, the mechanisms do not use any payments. All proofs, as well as additional examples, can be found in our full paper.

Obtaining Actual Truthfulness. Recall that in [4], the authors added a "reaction stage" to the standard mechanism design model in order to achieve actual truthfulness from approximate truthfulness (which is obtained via differential privacy). We can also use this model and their techniques to obtain actual truthfulness in our results. In our full paper, we also describe an alternative model where actual truthfulness can be obtained from approximate truthfulness. In this new model, the mechanism is given the ability to verify the truthfulness of a small number of players. This model is simple to use and is realistic in settings where the truthfulness of a player can be verified objectively (e.g., income, expenses, age, address).

2 Preliminaries and Definitions

For any $k \in \mathbb{N}$, we will use $[k]$ to denote the set $\{1, \ldots, k\}$. We consider a standard mechanism design environment consisting of the following components:

- A number n of *players*; we will often use $[n]$ to denote the set of n players.
- A *type space* T; each player has a private type from the type space T.

- A distribution $\mathcal{T}$ over the type space; the players' private types are independently drawn from this distribution.
- A set S of *social alternatives*; for convenience, we assume that S is finite.
- For each player i, a *utility function* $u_i : T \times S \to \mathbb{R}$; for $t \in T$ and $s \in S$, $u_i(t, s)$ represents the utility that player i receives if player i has type t and the social alternative s is chosen.

We will focus on direct revelation mechanisms where each player reports his/her type. Therefore, a *mechanism* is a function $M : T^n \to S$, and a (pure) strategy for player i is a function $\sigma_i : T \to T$ that maps true types to announced types. For convenience, whenever we refer to a mechanism $M : T^n \to S$, we assume that it is associated with an environment as described above.

2.1 Equilibrium Concepts

In this section, we will define several equilibrium concepts based on the standard Bayes-Nash equilibrium (see, e.g., [9]). These equilibrium concepts will be used to define various notions of truthfulness. Our definitions build on the *ϵ-Bayes-Nash equilibrium*, which is a relaxation of the Bayes-Nash equilibrium in the sense that an ϵ margin is allowed in the Nash conditions. This relaxation reflects the assumption that players will not deviate from the equilibrium if gains from deviation are sufficiently small. In this paper, we also refer to ϵ-Bayes-Nash equilibria as approximate Bayes-Nash equilibria. For more information about various notions of approximate equilibria, see [10–12].

Our equilibrium concepts strengthen the ϵ-Bayes-Nash equilibrium. We chose two strengthenings to address the following weaknesses of Nash equilibria. Firstly, a player's part of a Nash equilibrium is only guaranteed to be a best-response if all the other players are playing their parts of the equilibrium. In other words, a Nash equilibrium cannot tolerate players deviating from their equilibrium strategy — if there is one irrational person in the system, the equilibrium breaks down. Deviations are especially problematic in ϵ-equilibria, where there is less confidence that everyone would play their part of the equilibrium. Secondly, a Nash equilibrium is not resilient to deviations by more than one person; coalitions of players can have profitable deviations from the equilibrium.

To address the first problem, we strengthen the Nash conditions such that even if up to k players deviate from the equilibrium, everyone else's best-response is still to adhere to their part of the equilibrium. In other words, the equilibrium *tolerates* arbitrary deviations of k individuals.

Definition 1 (k-tolerant ϵ-Bayes-Nash equilibrium). *A strategy profile $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is a k-*tolerant ϵ-Bayes-Nash equilibrium *if for every $I \subseteq [n]$ with $|I| \leq k$, every possible announced types $\boldsymbol{t}'_I \in T^{|I|}$ for I, every player $i \notin I$, and every pair of types t_i, t'_i for player i, we have*

$$\mathbb{E}_{\boldsymbol{t}_J}[u_i(t_i, M(\sigma_i(t_i), \boldsymbol{t}'_I, \boldsymbol{\sigma}_J(\boldsymbol{t}_J)))] \geq \mathbb{E}_{\boldsymbol{t}_J}[u_i(t_i, M(t'_i, \boldsymbol{t}'_I, \boldsymbol{\sigma}_J(\boldsymbol{t}_J)))] - \epsilon,$$

where $J = [n] \setminus (I \cup \{i\})$ and $\boldsymbol{t}_J \sim \mathcal{T}^{|J|}$.

We note that k-tolerance is distinct from the notion of *k-immunity* as defined in [13, 14], which guarantees that when up to k people deviate from the equilibrium, the utilities of the non-deviating players do not decrease.

The second problem mentioned above is addressed by *r-resilience* (see, e.g., [10, 14]). A Bayes-Nash equilibrium is r-resilient if for any group of size at most r, there does not exist a deviation of the group such that any member of the coalition has increased utility.

Definition 2 (r-resilient ϵ-Bayes-Nash equilibrium). *A strategy profile $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is an* r-resilient ϵ-Bayes-Nash equilibrium *if for every coalition $C \subseteq [n]$ with $|C| \leq r$, every true types $\boldsymbol{t}_C \in T^{|C|}$ for C, every player $i \in C$, and every possible announced types $\boldsymbol{t}'_C \in T^{|C|}$ for C, we have*

$$\mathbb{E}_{\boldsymbol{t}_{-C}}[u_i(t_i, M(\boldsymbol{\sigma}_C(\boldsymbol{t}_C), \boldsymbol{\sigma}_{-C}(\boldsymbol{t}_{-C})))] \geq \mathbb{E}_{\boldsymbol{t}_{-C}}[u_i(t_i, M(\boldsymbol{t}'_C, \boldsymbol{\sigma}_{-C}(\boldsymbol{t}_{-C})))] - \epsilon,$$

where $\boldsymbol{t}_{-C} \sim \mathcal{T}^{n-|C|}$.

It is not hard to see that resilience and tolerance can be independently violated, and hence neither implies the other. Just as the authors in [13, 14] combine immunity and resilience, we consider the combination of tolerance and resilience. Roughly speaking, a *(k, r)-persistent Bayes-Nash equilibrium* is a Bayes-Nash equilibrium that is r-resilient (protects against coalitions of size r), even in the presence of up to k individuals that are deviating arbitrarily from the equilibrium.

Definition 3 ((k, r)-persistent ϵ-Bayes-Nash equilibrium). *A strategy profile $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is a* (k, r)-persistent ϵ-Bayes-Nash equilibrium *if for every $I \subseteq [n]$ with $|I| \leq k$, every possible announced types $\boldsymbol{t}'_I \in T^{|I|}$ for I, every coalition $C \subseteq [n] \setminus I$ with $|C| \leq r$, every true types $\boldsymbol{t}_C \in T^{|C|}$ for C, every player $i \in C$, and every possible announced types $\boldsymbol{t}'_C \in T^{|C|}$ for C, we have*

$$\mathbb{E}_{\boldsymbol{t}_J}[u_i(t_i, M(\boldsymbol{\sigma}_C(\boldsymbol{t}_C), \boldsymbol{t}'_I, \boldsymbol{\sigma}_J(\boldsymbol{t}_J)))] \geq \mathbb{E}_{\boldsymbol{t}_J}[u_i(t_i, M(\boldsymbol{t}'_C, \boldsymbol{t}'_I, \boldsymbol{\sigma}_J(\boldsymbol{t}_J)))] - \epsilon,$$

where $J = [n] \setminus (I \cup C)$ and $\boldsymbol{t}_J \sim \mathcal{T}^{|J|}$.

2.2 Notions of Truthfulness

In this section, we define various notions of truthfulness based on the equilibrium concepts from the previous section. Recall that a mechanism is *Bayes-Nash truthful* if the truthful strategy profile is a Bayes-Nash equilibrium. Similarly, a mechanism is *ϵ-Bayes-Nash truthful* if the truthful strategy profile is an ϵ-Bayes-Nash equilibrium. By using the equilibrium concepts from the previous section, we can obtain stronger notions of truthfulness.

Definition 4 ([k-tolerant]/[r-resilient]/[(k, r)-persistent] ϵ-Bayes-Nash truthful). *A mechanism is* k-tolerant ϵ-Bayes-Nash truthful *if the truthful strategy profile is a k-tolerant ϵ-Bayes-Nash equilibrium. Similarly, a mechanism is* r-resilient *(resp.,* (k, r)-persistent*)* ϵ-Bayes-Nash truthful *if the truthful strategy profile is an r-resilient (resp., (k, r)-persistent) ϵ-Bayes-Nash equilibrium.*

It is easy to see that if a mechanism is (k, r)-persistent ϵ-Bayes-Nash truthful, then it is also k-tolerant ϵ-Bayes-Nash truthful and r-resilient ϵ-Bayes-Nash truthful. In many settings, it is reasonable to believe that players in an ϵ-Bayes-Nash truthful mechanism will be truthful, since (1) truth-telling is simple while computing a profitable deviation can be costly (see, e.g., [15]), and (2) lying can induce a psychological (morality) cost. Indeed, there are many results in mechanism design that assume that approximate truthfulness is enough to ensure that players will be truthful (see, e.g., [1, 16–18]).

3 Privacy for Bayesian Mechanism Design

In this section, we describe and define *Bayesian differential privacy*, which is a natural adaptation of *differential privacy* [7, 8] to the Bayesian mechanism design setting. Roughly speaking, differential privacy requires that when one person's input to the mechanism is changed, the output distribution of the mechanism changes very little (here, the mechanism is randomized).

We now describe Bayesian differential privacy. We first note that even though the players' true types are drawn from some distribution $\mathcal{T}$, if all the players are non-truthful and announce a type independently of their true type, then the input of the mechanism is no longer distributional and we are essentially in the same scenario as in (worst-case) differential privacy. Thus, it is necessary to make some assumptions on the strategies of the players, so that the input of the mechanism contains at least some randomness.

In our notion of Bayesian differential privacy, we assume that at least some players (e.g., a constant fraction of the players) are truthful so that their announced types have the same distribution as their true types. This assumption is not unreasonable, since we later show that if a mechanism is Bayesian differentially private, then the mechanism is automatically persistent approximately truthful, so we expect that most players would be truthful anyway. In particular, if we have an equilibrium where most players are truthful, then privacy is achieved at this equilibrium.

Roughly speaking, (k, ϵ, δ)-*Bayesian differentially privacy* requires that when a player i changes his/her announced type, the output distribution of the mechanism changes by at most an (ϵ, δ) amount, assuming that at most k players are non-truthful (possibly lying in an arbitrary way). This implies that the mechanism leaks very little information about each player's announced type, so each player's privacy is protected. The mechanism is assumed to be deterministic, so the randomness of the output is from the randomness of the types of the truthful players. (One can also consider randomized mechanisms, but we chose to focus on deterministic mechanisms in this paper.)

Definition 5 ((k, ϵ, δ)-Bayesian differential privacy). *A mechanism $M : T^n \to S$ is* (k, ϵ, δ)-Bayesian differentially private *if for every player $i \in [n]$, every subset $I \subseteq [n] \setminus \{i\}$ of players with $|I| \leq k$, every pair of types $t_i, t_i' \in T$ for player i, and every $\boldsymbol{t}_I' \in T^{|I|}$, the following holds: Let $J = [n] \setminus (I \cup \{i\})$ (the*

remaining players) and $\boldsymbol{t}_J \sim \mathcal{T}^{|J|}$*; then, for every* $Y \subseteq S$*, we have*

$$\Pr[M(t_i, \boldsymbol{t}'_I, \boldsymbol{t}_J) \in Y] \leq e^{\epsilon} \cdot \Pr[M(t'_i, \boldsymbol{t}'_I, \boldsymbol{t}_J) \in Y] + \delta,$$

where the probabilities are over $\boldsymbol{t}_J \sim \mathcal{T}^{|J|}$*.*

The parameter k controls how many non-truthful players the mechanism can tolerate while satisfying privacy; k can be a function of n (the number of players), such as $k = \frac{n}{2}$. One can even view the non-truthful players as being controlled/known by an adversary that is trying to learn information about a player i's type; as long as the adversary controls/knows at most k people, player i's privacy is still protected. The parameters ϵ and δ bound the amount of information about each person's (announced) type that can be "leaked" by the mechanism. Since the above definition of Bayesian differential privacy is a natural adaptation of differential privacy to Bayesian mechanism design, and since differential privacy is a well-motivated and well-accepted notion of privacy, we will not further elaborate on the details of the above definition.

Our definition of (k, ϵ, δ)-Bayesian differential privacy has some similarities to the notion of (ϵ, δ)-*noiseless privacy* (for databases) introduced and studied in [19]. However, there are some subtle but significant differences between the two definitions, so the results in this paper do not follow from the theorems and proofs in [19]. Nevertheless, the ideas and techniques in [19], and for (ϵ, δ)-noiseless privacy in general, may be useful for designing Bayesian differentially private mechanisms.

It is known that differentially private mechanisms are approximately (dominant-strategy) truthful (see [1]). Similarly, Bayesian differentially private mechanisms are persistent approximate Bayes-Nash truthful.

Theorem 1 (Bayesian differential privacy $\implies$ persistent approximate truthfulness). *Suppose the utility functions are bounded by* $\alpha > 0$*, i.e., the utility function for each player* i *is* $u_i : T \times S \to [-\alpha, \alpha]$*. Let* M *be any mechanism that is* (k, ϵ, δ)*-Bayesian differentially private. Then,* M *satisfies the following properties:*

1. M *is* k*-tolerant* $(\epsilon + 2\delta)(2\alpha)$*-Bayes-Nash truthful.*
2. *For every* $1 \leq r \leq k+1$*,* M *is* r*-resilient* $(r\epsilon + 2r\delta)(2\alpha)$*-Bayes-Nash truthful.*
3. *For every* $1 \leq r \leq k+1$*,* M *is* $(k - r + 1, r)$*-persistent* $(r\epsilon + 2r\delta)(2\alpha)$*-Bayes-Nash truthful.*

4 Efficient Bayesian Mechanisms with Privacy and Persistent Approximate Truthfulness

In this section, we present three classes of mechanisms that achieve both Bayesian differential privacy and persistent approximate truthfulness.

4.1 Histogram Mechanisms

We first present a broad class of mechanisms, called *histogram mechanisms*, that achieve Bayesian differential privacy and persistent approximate truthfulness. Given a partition $P = \{B_1, \ldots, B_m\}$ of the type space T with m blocks (ordered in some way), a *histogram* with respect to P is simply a vector in $(\mathbb{Z}_{\geq 0})^m$ that specifies a count for each block of the partition. Given a partition P, let $\mathcal{H}_P$ denote the set of all histograms with respect to P; given a vector $\boldsymbol{t}$ of types, let $H_P(\boldsymbol{t})$ be the histogram formed from $\boldsymbol{t}$ by simply counting how many components (types) of $\boldsymbol{t}$ belong to each block of the partition P.

We now define what we mean by a *histogram mechanism*. Intuitively, a histogram mechanism is a mechanism that, on input a vector of types, computes the histogram from the types with respect to some partition P, and then applies any function $f : \mathcal{H}_P \to S$ to the histogram to choose a social alternative in S.

Definition 6 (Histogram mechanism). *Let P be any partition of the type space T. A mechanism $M : T^n \to S$ is a* histogram mechanism with respect to *P if there exists a function $f : \mathcal{H}_P \to S$ such that $M(\boldsymbol{t}) = f(H_P(\boldsymbol{t}))\ \forall\ \boldsymbol{t} \in T^n$.*

The following theorem states that any histogram mechanism with bounded utility functions and positive expected count for each bin is both Bayesian differentially private and persistent approximately truthful.

Theorem 2 (Histogram mechanisms are private and persistent approximately truthful). *Let $M : T^n \to S$ be any histogram mechanism with respect to some partition P of T. Let $p_{\min} = \min_{B \in P} \Pr_{t \sim \mathcal{T}}[t \in B]$, and suppose that $p_{\min} > 0$. Then, for every $0 \leq k \leq n-2$ and $\frac{4}{p_{\min} \cdot (n-k-1)} \leq \epsilon \leq 1$, M satisfies the following properties with $\delta = e^{-\Omega((n-k) \cdot p_{\min} \cdot \epsilon^2)}$:*

1. Privacy: *M is (k, ϵ, δ)-Bayesian differentially private.*
2. Persistent approximate truthfulness: *Suppose the utility functions are bounded by $\alpha > 0$, i.e., the utility function for each player i is $u_i : T \times S \to [-\alpha, \alpha]$. Then, for every $1 \leq r \leq k+1$, M is $(k-r+1, r)$-persistent $(r\epsilon + 2r\delta)(2\alpha)$-Bayes-Nash truthful.*

One possible partition of the type space is the one where there is a distinct block for each type. Thus, Theorem 2 covers the case where the choice of the mechanism depends only on the *number* of players that reported each type, and not their identities. In fact, given any partition, one can redefine the type space so that the new types are the blocks of the partition. This means we could always redefine the type space and simply use the partition where there is a distinct block for each type in the new type space. However, we believe it is more natural to preserve the original, natural type space, and to allow the histogram mechanism to use an appropriate partition of the type space.

In Theorem 2, since the histogram mechanism is not modified in any way to satisfy privacy and persistent approximate truthfulness, all properties of the mechanism (e.g., efficiency, truthfulness, individual rationality, etc.) are preserved. We now give a simple example to illustrate Theorem 2.

*Example 1 (***Voting with multiple candidates).** Suppose we are trying to select a winner from a finite set of candidates (e.g., political candidates) using the plurality rule (i.e., each voter casts one vote and the candidate with the most votes wins). The set of social alternatives is the set of candidates, and the natural type space is the set of all preference orders over the candidates. However, we can partition the type space such that each block b represents a candidate c_b, and all the types with c_b as their top choice belong to block b. Intuitively, announcing a type that belongs to block b can be understood as casting a vote for candidate c_b. Using this partition, we can define a histogram mechanism that implements the plurality rule. It is well known that the plurality rule is not strategy-proof when there are more than two candidates (see, e.g., [11]). However, by Theorem 2, this histogram mechanism is Bayesian differentially private and persistent approximate Bayes-Nash truthful.

4.2 Mechanisms for Two Social Alternatives

Although histogram mechanisms are useful in many settings, in order to apply Theorem 2 to get good parameters, the number of bins cannot be extremely large. We now present a class of mechanisms that do not require the partitioning of types into bins, but are still Bayesian differentially private and persistent approximately truthful. Roughly speaking, the following theorem states that any mechanism that makes a choice between two social alternatives $\{A, B\}$ based on the difference between the sums of two functions $u(\cdot, A)$ and $u(\cdot, B)$ on the types is Bayesian differentially private and persistent approximately truthful.

Theorem 3 (Private and persistent approximately truthful mechanisms for two social alternatives). *Let $S = \{A, B\}$ be any set of two social alternatives, let $T \subseteq \mathbb{R}$ be the type space, let $\mathcal{T}$ be any distribution over T, and let $u : T \times S \to [-\beta, \beta]$ be any function (e.g., a utility function for all the players). Suppose the random variable $u(t, A) - u(t, B)$, where $t \sim \mathcal{T}$, has non-zero variance and a probability density function.*

Let $M : T^n \to S$ be any mechanism such that

$$M(\boldsymbol{t}) = f\left(\sum_{i=1}^{n} u(t_i, A) - \sum_{i=1}^{n} u(t_i, B)\right)$$

for some function $f : \mathbb{R} \to S$. Then, for every $0 \leq k \leq n-2$ and $0 < \epsilon \leq 1$, M satisfies the following properties with $\epsilon' = \epsilon + O(\sqrt{\frac{\ln(n-k)}{n-k}})$ and $\delta = O(\frac{1}{\epsilon\sqrt{n-k}})$:

1. Privacy: *M is (k, ϵ', δ)-Bayesian differentially private.*
2. Persistent approximate truthfulness: *Suppose the utility functions are bounded by $\alpha > 0$, i.e., the utility function for each player i is $u_i : T \times S \to [-\alpha, \alpha]$. Then, for every $1 \leq r \leq k+1$, M is $(k-r+1, r)$-persistent $(r\epsilon' + 2r\delta)(2\alpha)$-Bayes-Nash truthful.*

The mechanism in Theorem 3 chooses a social alternative by applying some function f on the difference between $\sum_{i=1}^{n} u(t_i, A)$ and $\sum_{i=1}^{n} u(t_i, B)$. We note that the mechanism may already have certain properties, such as efficiency, truthfulness, individual rationality, etc.; by Theorem 3, this mechanism also satisfies privacy and persistent approximate truthfulness, in addition to the original properties that it already satisfies. One obvious application of Theorem 3 is to let u be a common utility function for the players, where the utility of player i with type t_i is $u(t_i, A)$ if A is chosen, and is $u(t_i, B)$ if B is chosen. If we define $f : \mathbb{R} \to S$ such that $f(x) = A$ if and only if $x > 0$, then the mechanism maximizes social welfare.

4.3 Social Welfare Mechanisms

In this section, we present a class of mechanisms that make choices based on the social welfare provided by each social alternative. An important subset of these mechanisms is the set of mechanisms that maximize social welfare.

In this section, a type $t \in T$ is a valuation function that assigns a valuation to each social alternative $s \in S$. In many settings, it is reasonable to assume that the players' valuations for each social alternative follow a normal distribution, since the normal distribution has been frequently used to model many natural and social phenomena. For convenience of presentation, we will use the *standard* normal distribution $\mathcal{N}(0, 1)$ in our theorems below. However, our theorems can be easily generalized to work with arbitrary normal distributions. In any case, it is easy to see that given any normal distribution over the valuations, the valuations can be translated and scaled to obtain the standard normal distribution.

For any reasonable mechanism, it is natural to have a bound on the set of possible valuations — it would be unreasonable to allow a player to report an arbitrarily high or low valuation (e.g. 2^{100}) and single-handedly influence the choice of the mechanism. Therefore, we will restrict the possible valuations to the interval $[-\alpha, \alpha]$ for some value $\alpha > 0$. As a result, our type space T will be the set of all valuation functions $t : S \to [-\alpha, \alpha]$. Furthermore, we will assume that the players' valuations for each social alternative follow the standard normal distribution. However, because of the bound on the set of valuations, we will use the truncated standard normal distribution obtained by conditioning $\mathcal{N}(0, 1)$ to lie on the interval $[-\alpha, \alpha]$. We denote this distribution by $\mathcal{N}(0, 1)_{[-\alpha,\alpha]}$.

For simplicity, we will first present the following theorem, which is a special case of our more general result (Theorem 5). The following theorem states that if each player's valuation for each social alternative is distributed as the truncated standard normal distribution $\mathcal{N}(0, 1)_{[-\alpha,\alpha]}$, then any mechanism that makes a choice based on the set of total valuations for each social alternative is Bayesian differentially private and persistent approximate Bayes-Nash truthful.

Theorem 4 (Social welfare mechanisms). *Let $S = \{s_1, \ldots, s_m\}$ be a set of m social alternatives. Let the type space T be the set of all valuation functions $t : S \to [-\alpha, \alpha]$ on S, where $\alpha = \Theta(\sqrt[4]{n})$. Let $\mathcal{T}$ be the distribution over T*

obtained by letting $t(s) \sim \mathcal{N}(0,1)_{[-\alpha,\alpha]}$ *for each* $s \in S$ *independently. For each player* i*, let the utility function for player* i *be* $u_i(t_i, s) = t_i(s)$*.*

Let $\mathsf{sw}_j(\boldsymbol{t}) = \sum_{i=1}^{n} t_i(s_j)$ *be the (reported) social welfare for the social alternative* s_j*. Let* $M : T^n \to S$ *be any mechanism such that*

$$M(\boldsymbol{t}) = f(\mathsf{sw}_1(\boldsymbol{t}), \ldots, \mathsf{sw}_m(\boldsymbol{t}))$$

for some function $f : \mathbb{R}^m \to S$*. Then, for every constant* $c < 1$*, every* $k \leq c \cdot n$*, and every* $0 < \epsilon \leq 1$*,* M *satisfies the following properties with* $\delta = O(e^{-\Omega(\frac{\epsilon^2}{m^2} \cdot \sqrt{n}) + \ln(m\sqrt{n})})$*:*

1. Privacy: M *is* (k, ϵ, δ)*-Bayesian differentially private.*
2. Persistent approximate truthfulness: *For every* $1 \leq r \leq k+1$*,* M *is* $(k - r + 1, r)$*-persistent* $(r\epsilon + 2r\delta)(2\alpha)$*-Bayes-Nash truthful.*

In Theorem 4, $\mathsf{sw}_j(\boldsymbol{t})$ represents the social welfare that will be achieved if the players' types (i.e., valuation functions) are $\boldsymbol{t}$ and the social alternative s_j is chosen by the mechanism. Thus, Theorem 4 says that any mechanism whose choice depends only on the set $\{\mathsf{sw}_j(\boldsymbol{t})\}_{j \in [m]}$ of social welfare values satisfies Bayesian differential privacy and persistent approximate Bayes-Nash truthfulness, in addition to any properties that it may already satisfy (e.g., efficiency, truthfulness, individual rationality, etc.). In particular, a mechanism that chooses a social alternative to maximize social welfare satisfies this requirement and achieves optimal efficiency with respect to social welfare.

In Theorem 4, the value α at which the standard normal distribution is truncated is chosen so that the truncated distribution is very close to the untruncated one. This means that even if we had used the untruncated distribution instead, with high probability no valuation would fall outside the interval $[-\alpha, \alpha]$.

In the next theorem, we consider a setting where there is a set of available "options", and we allow the mechanism to choose any subset of these options. Thus, the set of social alternatives is the power set of the set of options. To keep the set of valuations tractable, instead of having a valuation for each social alternative, the players have a valuation for each option. Moreover, we allow for the flexibility where for each player, only certain options are relevant/applicable to that player. We capture this flexibility by having a binary weight for each player-option pair. Note that Theorem 4 is the special case where the set of social alternatives consists of the sets of single options (i.e., the singletons), and where all options are considered relevant to all players.

The binary weight $w_{i,j}$ associated with player i and option o_j indicates whether option o_j is relevant/applicable to player i. $w_{i,j} = 1$ means that option o_j is relevant/applicable to player i, so player i's announced valuation is taken into account in the social welfare for option o_j. On the other hand, $w_{i,j} = 0$ means that player i's valuation is ignored in the social welfare for option o_j. These weights are known to or set by the mechanism designer. For example, perhaps only people with low income should have a voice in decisions regarding subsidized housing, and only people with disabilities should have a say in decisions regarding building accessibility laws. We now state our next theorem, which generalizes Theorem 4 to this new setting.

Theorem 5 (Social welfare mechanisms with multiple options). *Let the set S of social alternatives be 2^O, where $O = \{o_1, \ldots, o_m\}$ is a set of m possible "options". Let the type space T be the set of all valuation functions $t : O \to [-\alpha, \alpha]$ on O, where $\alpha = \Theta(\sqrt[4]{n})$. Let $\mathcal{T}$ be the distribution over T obtained by letting $t(o) \sim \mathcal{N}(0,1)_{[-\alpha,\alpha]}$ for each option $o \in O$ independently. Suppose the weights $\{w_{i,j}\}_{i\in[n],j\in[m]}$ satisfy $\sum_{i=1}^{n} w_{i,j} \geq c_1 \cdot n$ for every option o_j, where $c_1 > 0$ is some constant.*

Let $\mathsf{sw}_j(\boldsymbol{t}) = \sum_{i=1}^{n} w_{i,j} \cdot t_i(o_j)$ be the (reported) social welfare for option o_j. Let $M : T^n \to S$ be any mechanism such that

$$M(\boldsymbol{t}) = f(\mathsf{sw}_1(\boldsymbol{t}), \ldots, \mathsf{sw}_m(\boldsymbol{t}))$$

for some function $f : \mathbb{R}^m \to S$. Then, for every constant $c_2 < c_1$, every $k \leq c_2 \cdot n$, and every $0 < \epsilon \leq 1$, M satisfies the following properties with $\delta = O(e^{-\Omega(\frac{\epsilon^2}{m^2}\cdot\sqrt{n})+\ln(m\sqrt{n})})$:

1. Privacy: *M is (k, ϵ, δ)-Bayesian differentially private.*
2. Persistent approximate truthfulness: *Suppose the utility functions are bounded by $\beta > 0$, i.e., the utility function for each player i is $u_i : T \times S \to [-\beta, \beta]$. Then, for every $1 \leq r \leq k+1$, M is $(k-r+1, r)$-persistent $(r\epsilon + 2r\delta)(2\beta)$-Bayes-Nash truthful.*

In Theorem 5, the requirement on the binary weights simply means that each option is relevant/applicable to at least some constant fraction of the players. Note that the persistent approximate truthfulness result of Theorem 5 requires the players' utility functions to be bounded by $\beta > 0$. This assumption is needed since the players' utility functions can actually be arbitrary functions. However, the most natural way to use Theorem 5 is to let player i's utility function be the following: if the chosen social alternative is a singleton $\{o_j\}$, then the utility for player i is $w_{i,j} \cdot t_i(o_j)$; if the chosen social alternative is a set s consisting of two or more options, then the utility for player i is the sum of the utilities for each singleton subset of s. Alternatively, a player i's utility for a social alternative s does not have to be *additive* in the options that s contains — the utility function for player i can capture complementarities and substitutabilities of the options as well. We now give a simple example that illustrates Theorem 5.

*Example 2 (***Multiple public projects).** The municipal government would like to spend its budget surplus of 4 million on the community. There are four options that the government is considering, each costing 2 million to build: a senior home, a casino, a subsidized housing complex, and a library. The government would like to find out, on a scale from $-\alpha$ to α, how much each individual values each option. For each individual i, the government chooses the weights for each of the options as follows: the weight for the senior home is 1 if and only if individual i is over the age of 65; the weight for the casino is 1 if and only if individual i is over the age of 19; the weight for the subsidized housing complex is 1 if and only if individual i is classified as low-income; and the weight for the library is always 1.

After collecting the valuations from the individuals, the government can compute the social welfare provided by each option, or compute an average utility for each option by dividing its social welfare by the number of people who have weight 1 for that option. Finally, the government can choose two of the options to maximize social welfare or average utility. By Theorem 5, this mechanism is Bayesian differentially private and persistent approximately truthful.

In our full paper, we generalize Theorem 5 to the case where the players' valuations for each social alternative are arbitrarily distributed with non-zero variance.

References

1. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: FOCS. (2007) 94–103
2. Xiao, D.: Is privacy compatible with truthfulness? IACR ePrint (2011)
3. Chen, Y., Chong, S., Kash, I.A., Moran, T., Vadhan, S.P.: Truthful mechanisms for agents that value privacy. Manuscript (2011)
4. Nissim, K., Smorodinsky, R., Tennenholtz, M.: Approximately optimal mechanism design via differential privacy. In: ITCS. (2012) 203–213
5. Nissim, K., Orlandi, C., Smorodinsky, R.: Privacy-aware mechanism design. In: EC. (2012) 774–789
6. Huang, Z., Kannan, S.: The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In: FOCS. (2012)
7. Dwork, C., Mcsherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: TCC. (2006) 265–284
8. Dwork, C.: Differential privacy. In: ICALP. (2006) 1–12
9. Fudenberg, D., Tirole, J.: Game Theory. MIT Press (1991)
10. Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V.V.: Algorithmic Game Theory. Cambridge University Press (2007)
11. Shoham, Y., Leyton-Brown, K.: Multiagent Systems - Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press (2009)
12. Tijs, S.H.: Nash equilibria for noncooperative n-person games in normal form. SIAM Review **23**(2) (1981) pp. 225–237
13. Abraham, I., Dolev, D., Halpern, J., Gonen, R.: Distributed computing meets game theory: Robust mechanisms for rational secret sharing and multiparty computation. In: PODC. (2006) 53–62
14. Halpern, J.Y.: Beyond nash equilibrium: solution concepts for the 21st century. In: PODC. (2008)
15. Halpern, J., Pass., R.: Game theory with costly computation: Formulation and application to protocol security. In ICS (2010) 120–142
16. Birrell, E., Pass, R.: Approximately strategy-proof voting. In: IJCAI. (2011) 67–72
17. Kothari, A., Parkes, D.C., Suri, S.: Approximately-strategyproof and tractable multi-unit auctions (2004)
18. Schummer, J.: Almost-dominant strategy implementation: exchange economies. Games and Economic Behavior **48**(1) (2004) 154–170
19. Bhaskar, R., Bhowmick, A., Goyal, V., Laxman, S., Thakurta, A.: Noiseless database privacy. In: ASIACRYPT. (2011) 215–232