**THE PINHAS SAPIR CENTER FOR DEVELOPMENT**
**TEL AVIV UNIVERSITY**

# Privacy-Aware Mechanism Design*

Kobbi Nissim[i], Claudio Orlandi[ii], Rann Smorodinsky[iii]

Discussion Paper No. 4-14

March 2014

**The paper can be downloaded from: http://sapir.tau.ac.il**

Thanks to The Pinhas Sapir Center for Development at Tel Aviv University for their financial support.

A preliminary version appeared in ACM Conference on Electronic Commerce (EC), pp. 774{789A, 2012.

---

[i] Dept. of Computer Science, Ben-Gurion University of the Negev, Be'er Sheva, Israel. Supported by the European Research.Council as part of the ERC project SFEROT Sapir. kobbi@cs.bgu.ac.il.

[ii] Department of Computer Science, Aarhus University, Denmark. Work done while C.O. was visiting the Dept. of Computer Science, Bar-Ilan University, Ramat Gan, Israel. Supported by the European Research Council as part of the ERC project LAST. orlandi at cs.au.dk.

[iii] Fac. of Industrial Engineering and Management, Technion, Haifa, Israel. Supported by Technion VPR grants, TASP, the Bernard M. Gordon Center for Systems Engineering at the Technion. rann@ie.technion.ac.il

**Abstract**

Mechanism design deals with distributed algorithms that are executed with self-interested agents. The designer, whose objective is to optimize some function of the agents' private *types*, needs to construct a computation that takes into account agent incentives which are not necessarily in alignment with the objective of the mechanism. Traditionally, agents are modeled to only care about the utility they derive from the mechanism outcome, and mechanisms designed for such agents often fully or partially discloses the agents' declarations. When agents are *privacy aware*, i.e., their loss of privacy can adversely affects their utility, such mechanisms may become inadequate - ignoring privacy-awareness in the design of a mechanism may render it not incentive compatible, and hence inefficient. Interestingly, and somewhat counter-intuitively, Xiao [eprint 2011] has recently showed that this can happen even with mechanisms that preserve a strong notion of privacy.

Towards constructing mechanisms for privacy-aware agents, we put forward and justify a model of privacy-aware mechanism design. We then show that privacy-aware mechanisms are feasible. The following is a summary of our contributions:

- **Modeling privacy-aware agents:** We propose a new model of privacy-aware agents where agents need only have a conservative upper bound on how loss of privacy adversely affects their utility. This is in deviation from prior modeling which required a complete characterization.
- **Privacy of the privacy loss valuations:** Agent privacy valuations are often sensitive on their own. Our model of privacy-aware mechanisms takes into account the loss of utility due to information leaked about these valuations.
- **Guarantees for agents with high privacy valuations:** As it is impossible to guarantee incentive compatibility for agents that have arbitrarily high privacy valuations, we require a privacy-aware mechanism to set a threshold such that the mechanism is incentive compatible w.r.t.\ agents whose privacy valuations are below the threshold, and differential privacy is guaranteed for all other agents.
- **Constructing privacy-aware mechanisms:** We first construct a privacy-aware mechanism for a simple polling problem, and then give a more general result, based on recent generic construction of approximately additive mechanisms by Nissim, Smorodinsky, and Tennenholtz [ITCS 2012]. We show that under a mild assumption on the distribution of privacy valuations (namely, that valuations are bounded for all but a vanishing fraction of the population) these constructions are incentive compatible w.r.t.\ almost all agents, and hence give an approximation of the optimum. Finally, we show how to apply our generic construction to get a mechanism for privacy-aware selling of digital goods.

# 1 Introduction

Mechanism design deals with distributed algorithms that are executed with self-motivated agents who optimize their own objective functions. The mechanism designer, interested in computing some function of the agents' private inputs (henceforth types), needs to construct a computation that takes into account the agents' incentives, which are not necessarily in alignment with the goals of the designer. Settings where mechanism design is instrumental include centralized allocation of resources, pricing, deciding the level of provision of a public good, etc. Traditionally, agents are modeled to care about the utility they derive from the outcome of the mechanism, but not about their privacy. Consequently, in many cases the outcome of the mechanism discloses, fully or partially, the types declared by agents.

We look at a model where agents also assign non-positive utility to the leakage of information about their private types through the public outcome of the mechanism. This modeling is relevant, e.g., when private information is aggregated via markets which provide superior prediction power (e.g., [WZ04]), kidney exchange markets where information aggregation and sharing lead to huge health-care benefits (e.g., [AR11]), or recommendation engines which assist individuals in locating optimal products. Such markets may not be incentive compatible and consequently can fail if agents' privacy is not accounted for.

Our work is on the interface of the research in Algorithmic Game Theory and the recent theoretical research of privacy. Earlier scholarly work by McSherry and Talwar [MT07] has forged a link between the notion of *differential privacy* [DMNS06] and mechanism design. They observed that differential privacy can serve as a tool for constructing mechanisms where truthfulness is $\varepsilon$-dominant. A recent work [NST12] has observed a few weaknesses in constructions resulting from [MT07] and resolved them by putting forward a general framework for constructing approximately-optimal mechanisms where truthfulness is a dominant strategy or an ex-post Nash equilibrium. This line of work demonstrates that differential privacy can serve as a powerful *tool* for the construction of efficient mechanisms.

The mechanisms presented in [MT07, NST12] were not analyzed with respect to agents who take into account their dis-utility due to the information leaked about their types. We call this dis-utility *information utility* and we call those agents that take the information utility into account *privacy-aware*. It might be tempting to think that the combination of truthfulness and differential privacy is always sufficient for making privacy-aware agents truthful – mechanisms that are truthful and preserve differential privacy should remain truthful also with respect to agents that take information utility into account. A work of Xiao [Xia11] dispels this intuition by showing a mechanism that preserves differential privacy and is truthful with respect to agents that are not privacy aware, yet, under a particular definition of information utility, truthfulness is not dominant with respect to privacy-aware agents.

A recent work of Ghosh and Roth [GR11] constructs mechanisms that compensate agents for their loss in privacy. They consider a setting where a data analyst wishing to perform an $\varepsilon$-differentially private computation of a statistic pays the participating agents for using their data. They construct mechanisms where agents declare how their information utility depends on the privacy parameter $\varepsilon$, and the mechanism chooses $\varepsilon$, decides which agents' information will be used in the computation, and how much those agents will be paid. The mechanisms presented in [GR11] do not preserve the privacy of how each of the agents values her privacy. However, an agent's value for privacy can reveal information about the agents' private data: it is not unreasonable to assume that there is some correlation between the price an agents sets on her privacy and the unlikelihood of her private data or, in other words, to assume that people value their privacy more if they have something to hide.

In light of these issues, our goal is to construct mechanisms for privacy-aware agents, where privacy is accounted for the 'traditional' inputs to the mechanism (such as valuations, locations, etc.) but also, and for the first time to the best of our knowledge, with respect to the privacy valuation itself.

The results of [GR11] show, however, that this goal is too ambitious – no individually rational mechanism can compensate individuals for the information (dis)utility incurred due to information leaked about the privacy valuation from the public output unless the privacy valuations are bounded. To overcome this obstacle we focus on mechanisms for large populations of agents: We propose a relaxation where loss in privacy is accounted for all agents whose valuations are bounded, where the bound increases as the agent population grows. Hence, in large enough populations truthfulness is provided for all (or most of) the

agents. For the small fraction of agents who value their privacy too much for the mechanism to compensate, we provide $\varepsilon$-differential privacy with respect to whether their privacy valuations exceed the bound. The value of $\varepsilon$ improves (i.e., reduces) with the population size.

## 1.1 Contributions

**Modeling.** A major contribution of this work is a new notion of privacy awareness in mechanism design which we motivate in light of weaknesses found in prior modelings. In our model, privacy-aware agents hold a combined type consisting of their 'traditional' *game type* and their *privacy type*, where for the latter agents need only have a conservative upper bound on how loss of privacy adversely affects their utility. Agents care about leakage of information on both their game and privacy types. These features are an important difference with respect to previous work (e.g., [GR11]) where a full characterization of the information utility was required to achieve truthfulness, and furthermore, mechanisms did not take into account the information cost of the privacy type.

Note that if agents can have arbitrarily high privacy valuations, then it is impossible to a priori bound the information utility of a computation whose outcome depends on agents' private inputs, or, alternatively, on their choice whether to participate or not (see also a more elaborate argument in [GR11] in the specific context of truthful mechanisms for selling private information for statistical computations). To sidestep this inherent difficulty, we opt for a lesser requirement from a privacy-aware mechanism: the mechanism should set a threshold on the privacy valuation $v_{max}$ and a privacy parameter $\varepsilon$ such that the mechanism is incentive compatible w.r.t. agents whose privacy valuations are below $v_{max}$ and $\varepsilon$-differential privacy is guaranteed for all agents.

**Construction of Privacy-Aware Mechanisms.** We next demonstrate that privacy-aware mechanisms are feasible. Our first result illustrates some of our techniques: in Section 4.1, we provide a simple privacy-aware poll between two or more alternatives. The main idea is to make (traditional) dis-utility due to mis-reporting dominate the information utility, and hence preserve truthfulness. We set a bound $v_{max}$ on the privacy valuations, and treat agents differently according to whether their valuations are above $v_{max}$ or not:

- For agents whose privacy valuations are below the bound, the mechanism ensures that the agents are provided with a fair compensation for their privacy loss.
- For agents whose privacy valuations are too high for the mechanism to compensate, we provide that their privacy valuations are protected in a $\varepsilon$-differentially private way. As discussed above, this is in a sense the best we can hope to achieve.

We then move our attention to large populations and we introduce the notion of *admissible populations* by making a somewhat mild assumption on the distribution of the valuations (i.e., finiteness of its moments).

In Section 5 we present a generic construction of privacy-aware mechanisms. Our construction is based on the recent construction of [NST12] which we modify to accommodate privacy-agents sampled from an admissible population. We show that the mechanism achieves truthfulness for most agents and non-trivial accuracy. Finally in Section 5.2 we present a natural example of a privacy-aware mechanism that falls in our framework i.e., privacy-aware selling of digital goods.

In a sense, our results show that when the outcome of a truthful (not necessarily privacy-aware) mechanism is insensitive to each of its individual inputs (as is often the case when the underlying population is large), it is rational for most privacy-aware agents to report truthfully. This is because the information leaked about their private types is small, and hence bounded away from the decrease in utility that can be caused by misreporting their type.

## 1.2 Other Related Work

The cryptographic literature also includes references to "privacy-preserving mechanism design" (an example is the work of Naor, Pinkas and Sumner [NPS99]). We stress that our goals are different from these

cryptographic realizations of mechanisms as in our setting the agents are worried about what the public outcome of a mechanism may leak about their types and privacy valuations, whereas the goal of cryptographic realizations of mechanisms is to hide all information except for the outcome of the mechanism. As showed in [MNT09], using cryptography to implement mechanism designs over an Internet-like network is a non-trivial task, and one needs to make sure that the properties of the mechanism (e.g., truthfulness) are preserved also by the cryptographic implementation of the mechanism.

Independently from our work, Chen, Chong, Kash, Moran, and Vadhan [CCK+11] also studied the problem of truthful mechanisms in the presence of agents that value privacy. The motivation for both their work and ours is similar, and in both the quantification of privacy loss corresponds to the effect an agent's input has on the outcome of a mechanism. The model that [CCK+11] presents for privacy-aware agents (and hence privacy-aware mechanisms) is different from ours in that agents are assumed to value privacy on a *per-outcome* basis, whereas our modeling is more conservative as agents' privacy valuations depend on the *overall* (i.e., worst case) outcome of the mechanism. Both modelings are well motivated, our reliance on a weaker assumption may lead to more robust mechanisms, where the per-outcome approach may lead to a richer set of privacy-aware mechanisms.

# 2 Preliminaries

We refer to discrete sets $T$ and $S$ as the *type* set, and the set of *social alternatives* respectively. For two vectors $t, t' \in T^n$ we define the *Hamming distance* between $t$ and $t'$ to be the number of entries on which $t, t'$ differ, i.e., $|\{i : t_i \neq t'_i\}|$. Vectors that are within Hamming distance one are called *neighboring*.

A mechanism $M : T^n \to \Delta(S)$ is a function that assigns for any vector of inputs $t \in T^n$ a distribution over $S$ (the notation $\Delta(S)$ denotes the set of probability distributions over the set $S$). The outcome of an execution of $M$ on input $t \in T^n$ is an element $s \in S$ chosen according to the distribution $M(t)$, and the notation $M(t)(S)$ is used for the probability measure of $S$ w.r.t. probability distribution $M(t)$.

Differential privacy is a very conservative way of measuring how much information a mechanism leaks about the type of any given agent - even given *complete knowledge* of all other agent types. Formally:

**Definition 2.1** (Differential Privacy [DMNS06])**.** *A mechanism $M : T^n \to \Delta(S)$ preserves $\varepsilon$-differential privacy if for all neighboring $t, t' \in T^n$ and for all (measurable) subsets $S'$ of $S$ it holds that*

$$M(t)(S') \leq e^{\varepsilon} \cdot M(t')(S').$$

The following simple lemma follows directly from the above definition (the proofs for Lemma 2.2 and Theorem 2.4 below are not new and are included for completeness).

**Lemma 2.2.** *Let $M : T^n \to \Delta(S)$ be a mechanism that preserves $\varepsilon$-differential privacy and let $g : S \to \mathbb{R}^{\geq 0}$. Then, for all neighboring $t, t' \in T^n$*

$$\mathbf{E}_{s\sim M(t)}[g(s)] \leq e^{\varepsilon} \cdot \mathbf{E}_{s\sim M(t')}[g(s)].$$

*In particular, if $\varepsilon \leq 1$ and $g : S \to [0, 1]$,*

$$\left|\mathbf{E}_{s\sim M(t)}[g(s)] - \mathbf{E}_{s\sim M(t')}[g(s)]\right| < 2\varepsilon.$$

*Proof.* Let $t, t', g$ be as in the lemma.

$$\mathbf{E}_{s\sim M(t)}[g(s)] = \sum_{s\in S} M(t)(s)g(s) \leq \sum_{s\in S} e^{\varepsilon} M(t')(s)g(s) = e^{\varepsilon} \cdot \mathbf{E}_{s\sim M(t')}[g(s)],$$

where the inequality follows since $M$ provides $\varepsilon$-differential privacy, and $g$ is non-negative. For $\varepsilon \leq 1$ and $g : S \to [0, 1]$ we get

$$\mathbf{E}_{s\sim M(t)}[g(s)] - \mathbf{E}_{s\sim M(t')}[g(s)] \leq (e^{\varepsilon} - 1) \cdot \mathbf{E}_{s\sim M(t')}[g(s)] \leq e^{\varepsilon} - 1,$$

where the last inequality holds because $g$ returns a values in $[0,1]$. Similarly, we get $\mathbf{E}_{s\sim M(t')}[g(s)] - \mathbf{E}_{s\sim M(t)}[g(s)] \leq e^{\varepsilon} - 1$, hence

$$\left|\mathbf{E}_{s\sim M(t)}[g(s)] - \mathbf{E}_{s\sim M(t')}[g(s)]\right| \leq e^{\varepsilon} - 1 < 2\varepsilon,$$

where the last inequality follows noting that $(e^{\varepsilon} - 1) \leq 2\varepsilon$ for $0 \leq \varepsilon \leq 1$. □

A simple corollary of Lemma 2.2 is that $\left|\mathbf{E}_{s\sim M(t)}[\hat{g}(s)] - \mathbf{E}_{s\sim M(t')}[\hat{g}(s)]\right| < 4\varepsilon$ for neighboring $t, t'$ and $\hat{g} : S \to [-1,1]$.

**Definition 2.3** ([MT07])**.** *Let $f : S \times T^n \to \mathbb{R}^{\geq 0}$ and let $\varepsilon > 0$. The exponential mechanism for $f$ with parameter $\varepsilon$ is*

$$M_f^{\varepsilon}(t)(s) = \frac{\exp(\varepsilon f(s,t))}{\sum_{s'\in S}\exp(\varepsilon f(s',t))} \quad \text{for all } s \in S.$$

**Theorem 2.4** ([MT07])**.** *Let $\Delta f$ be the maximum of $f(s,t) - f(s,t')$ over all $s \in S$ and neighboring $t, t' \in T^n$. $M_f^{\frac{\varepsilon}{2\Delta f}}$ preserves $\varepsilon$-differential privacy.*

*Proof.* Let $t, t'$ be neighboring, and $S' \subseteq S$.

$$\begin{aligned}
M_f^{\frac{\varepsilon}{2\Delta f}}(t)(S') &= \sum_{s\in S'} \frac{\exp(\frac{\varepsilon}{2\Delta f} f(s,t))}{\sum_{s'\in S}\exp(\frac{\varepsilon}{2\Delta f} f(s',t))} \\
&= \sum_{s\in S'} \frac{\exp(\frac{\varepsilon}{2\Delta f}(f(s,t) - f(s,t')))\exp(\frac{\varepsilon}{2\Delta f} f(s,t'))}{\sum_{s'\in S}\exp(\frac{\varepsilon}{2\Delta f}((f(s',t) - f(s',t')))\exp(\frac{\varepsilon}{2\Delta f} f(s',t')))} \\
&\leq \sum_{s\in S'} \frac{\exp(\frac{\varepsilon}{2})\exp(\frac{\varepsilon}{2\Delta f} f(s,t'))}{\sum_{s'\in S}\exp(-\frac{\varepsilon}{2})\exp(\frac{\varepsilon}{2\Delta f} f(s',t')))} \\
&= \exp(\varepsilon) M_f^{\varepsilon}(t)(S'),
\end{aligned}$$

where the inequality follows by recalling that $\Delta f \geq |f(s,t) - f(s,t')|$ for all $s, t, t'$. □

**Definition 2.5** (Mutual Information)**.** *Let $X, Y$ be two random variables. The mutual information between $X$ and $Y$ is defined as*

$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$

*where $H(X) = -\sum_{x\in S}\Pr[X = x] \cdot \log(\Pr[X = x])$ is the Shannon entropy of $X$.*

It is well known that $I(X;Y) = H(X) - H(X|Y)$, i.e. $I(X;Y)$ measures the reduction in entropy in $X$ caused by conditioning on $Y$ (and symmetrically, $I(X;Y) = I(Y;X) = H(Y) - H(Y|X)$). The following simple observations follows from the Data Processing Inequality (see, e.g., [CT91, pp. 32]):

**Observation 2.6.** *For all (randomized) functions $f$,*

$$I(f(X);Y) \leq I(X;Y).$$

**Observation 2.7.** *Let $M : T^n \to \Delta(S)$ be an $\varepsilon$ differentially private mechanism, then for all random variables $X = (X_1, \ldots, X_n) \in \Delta(T^n)$ it holds that*

$$I(X_i; M(X), X_{-i}) \leq \varepsilon \log e.$$

*Proof.*

$$\begin{aligned}
H(M(X), X_{-i}|X_i) &= -\sum_{x_i \in T} p_{X_i}(x_i) \sum_{s \in S, x_{-i} \in T^{n-1}} p_{M(X),X_{-i}|X_i=x_i}(s, x_{-i}) \cdot \log p_{M(X),X_{-i}|X_i=x_i}(s, x_{-i}) \\
&\geq -\sum_{s \in S, x_{-i} \in T^{n-1}} \sum_{x_i \in T} p_{X_i}(x_i) \cdot p_{M(X),X_{-i}|X_i=x_i}(s, x_{-i}) \cdot \log \left( e^{\varepsilon} \cdot p_{M(X),X_{-i}}(s, x_{-i}) \right) \\
&= -\sum_{s \in S, x_{-i} \in T^{n-1}} p_{M(X),X_{-i}}(s, x_{-i}) \left( \varepsilon \log e + \log p_{M(X),X_{-i}}(s, x_{-i}) \right) \\
&= -\sum_{s \in S, x_{-i} \in T^{n-1}} p_{M(X),X_{-i}|X_i=x_i}(s, x_{-i}) \cdot \log p_{M(X),X_{-i}}(s, x_{-i}) - \varepsilon \log e \\
&= H(M(X), X_{-i}) - \varepsilon \log e,
\end{aligned}$$

where the inequality follows from the differential privacy of $M$. We conclude that

$$I(X_i; M(X), X_{-i}) = H(M(X), X_{-i}) - H(M(X), X_{-i}|X_i) \leq \varepsilon \log e.$$

□

# 3 Quantifying Information Utility

Our model is similar to the standard model of mechanism design, with the difference that agents participating in the execution of a mechanism care about their privacy. In the standard model, an agent's type $t_i$ expresses quantities such as a valuation of a good for sale, location, etc., the mechanism chooses an alternative $s$, and the agent's utility is a function of $t_i$ and $s$ (and sometimes, monetary transfers).

When considering privacy-aware agents, we need to introduce the information utility into their utility functions. A first issue that emerges is *how should this dis-utility be quantified?* Note that as different agents may value privacy differently, the quantification should be parametrized by agents' privacy preferences. We denote by $v_i$ the privacy preference of agent $i$. That is, an agent type is now composed of the 'traditional' type $t_i$, and a privacy preference $v_i$. A second issue that now emerges is that the alternative chosen by the mechanism can leak information about both $t_i$ and $v_i$, and hence leakage about $v_i$ needs also be taken into account.

## 3.1 Prior Work

In an early work, McGrew, Porter, and Shoham [MPS03] introduced privacy into agents' utility in the context of non-cooperative computing (NCC) [ST05]. In their model, agents only care about the case where other agents learn their private types with certainty. This means that privacy is either completely preserved or completely breached, and hence information utility is quantified to be either zero (no breach) or an agent dependent value $v_i > 0$. As it is often the case that leaked information is partial or uncertain, we are interested in more refined measures that take partial exposure into account.

A recent work by Ghosh and Roth [GR11] considers a setting where a data analyst wishes to perform a computation that preserves $\varepsilon$-differential privacy and compensates participating agents for their privacy loss. They assume a model where each agent's dis-utility is proportional to the privacy parameter $\varepsilon$. I.e., the $i$th agent's dis-utility is

$$u_i^{\text{inf}} = v_i \cdot \varepsilon,$$

where $v_i \geq 0$ is part of the agent's private type. A problem with this quantification is that while $\varepsilon$ measures the worst effect the $\varepsilon$-differentially private computation can have on privacy, the typical effect on agent $i$ can be significantly lower (see [DRV10]). In other words, while a differentially private mechanism guarantees that even in the worst-case one's privacy is not affected more than $\varepsilon$, a much better privacy may be provided *on average*. Furthermore, the actual information loss can depend on the other agents' inputs to the computation.

Another problem, that will be further discussed later, is that this quantification does not consider the information utility due to leakage of information about $v_i$ itself.

The third example we are aware of is from another recent work, by Xiao [Xia11]. Similarly to the present work, Xiao considers the setting of mechanism design with privacy-aware agents. The information utility is modeled to be

$$u_i^{\inf} = v_i \cdot I(t_i; M(t_{-i}, \sigma(t_i))),$$

where $v_i \geq 0$ is the agent privacy valuation. Note that with this measure, the dis-utility of agent $i$ depends on the distribution of her and the other agents' types, and on her own strategy $\sigma$. The following example demonstrates that this dependency on $\sigma$ can be problematic.

Consider the single-agent mechanism below, where the agent's private type consists of a single bit:

**Example 3.1** (The "Rye or Wholewheat" game)**.** *Alice is preparing a sandwich for Bob and inquires whether he prefers Rye (R) or Wholewheat (W). Bob wants to enjoy his favorite sandwich, but does not want Alice to learn his preference. Assume that Bob's type is uniformly chosen in* $\{\mathrm{R}, \mathrm{W}\}$ *and consider these two possibilities for Bob's strategy:*

1. *If Bob provides his true preference he will enjoy the sandwich. However, his information (dis)utility would be maximized as* $I(t_{Bob}; M(\sigma_{truthful}(t_{Bob}))) = 1$.
2. *If Bob answers at random he will enjoy the sandwich with probability one-half.*[1] *However, as his response does not depend on his preference no loss in privacy would be incurred, hence we get*

$$I(t_{Bob}; M(\sigma_{random}(t_{Bob}))) = 0 \ .$$

Note that since Bob's type is R or W with equal probability Alice's views of Bob's actions (and hence also the outcome of the mechanism) are distributed identically whether he uses $\sigma_{truthful}$ or $\sigma_{random}$. Hence, while mutual information differs dramatically between the two strategies – suggesting that Bob is suffering a privacy loss due to Alice learning his type in one but not in the other – it is impossible for Alice to distinguish between the two cases!

A few more words are in place regarding the source of this problem. First note that while the example demonstrates that $I(t_{Bob}; M(\sigma(t_{Bob})))$ is problematic as a measure of privacy it does not imply that $\sigma_{random}$ (nor the one time pad) is at fault (in fact, $\sigma_{random}$ provides Bob with perfect privacy even in a setting where Alice gets to know which strategy Bob uses, a guarantee $\sigma_{truthful}$ definitely does not provide). What the example capitalizes on is the fact that under the standard game-theoretic modeling it is possible that Alice does not get to see what Bob's strategy is. In such situations, it can happen that the more robust $\sigma_{random}$ is an overkill, as it provides Bob with less utility. We hence argue that the notion of information cost should be free of making assumptions on Alice's knowledge of $\sigma$.

## 3.2 Our Approach

We deviate from the works cited above as we do not present a new *exact* measure for information utility. We use a significantly weaker notion instead. To motivate our approach, re-consider the measures discussed above.

Looking first at the measure in [GR11], i.e., $v_i \cdot \varepsilon$, we note that while in $\varepsilon$-differential mechanisms the ratio $\Pr[M(t) = s] / \Pr[M(t') = s]$ is bounded by $e^{\varepsilon}$ for all neighboring $t, t'$ and $s$, it is plausible that the worst case behavior (i.e., outputting $s$ such that $\Pr[M(t) = s] / \Pr[M(t') = s] = e^{\varepsilon}$) occurs with only a tiny probability. This suggests that while $v_i \cdot \varepsilon$ may not be a good measure for information utility, it can serve as a good *upper bound* for this utility. Examining the measure in [Xia11] and trying to avoid the problem demonstrated in Example 3.1 above, we note that by Observation 2.6 $I(t_i; M(t)) \geq I(t_i; M(t_{-i}, \sigma(t_i)))$ for all $\sigma$, hence, we get that $v_i \cdot I(t_i; M(t))$ is another plausible *upper bound* for information utility. Finally, taking into account Observation 2.7 we get that $I(t_i; M(t)) \leq \varepsilon \log e$ and hence we choose to use $v_i \cdot \varepsilon$ as it is the weaker of these bounds.

[1] This is equivalent to encrypting Bob's type using a *one time pad*.

**Note 3.2.** *We emphasize that although our usage of the term $v_i \cdot \varepsilon$ is syntactically similar to that of [GR11], our usage of this quantity is conceptually very different. In particular, while loss of privacy cannot be used in our constructions for deterring non-truthful agents, the constructions (and proofs) in [GR11] use the fact that the information utility is (at least) $v_i \cdot \varepsilon$ for arguing truthfulness.*

**Note 3.3.** *Lemma 2.2 supports using $v_i \cdot \varepsilon$ as an upperbound for information utility in the follwing sense. An individual's concern about her privacy corresponds to a potential decrease in* future *utility due to information learned about her. An upper bound on information utility hence should correspond to this (potential) loss in future utility. By Lemma 2.2, the information contributed by individual $i$ affects the expectation of every non-negative (similarly, non-positive) function $g$ by* at most *a factor of $e^{\varepsilon}$. Let $G_i : S \to \mathbb{R}$ describe how the future utility of individual $i$ depends on the outcome of $M$. By Lemma 2.2, the information utility of that individual is bounded by*

$$\max_{t \in T^n} (e^{\varepsilon} - 1) \cdot \mathbf{E}_{s \sim M(t)} \left| G_i(s) \right| \approx \varepsilon \cdot \max_{t \in T^n} \mathbf{E}_{s \sim M(t)} \left| G_i(s) \right|,$$

*where the approximation holds for small $\varepsilon$. See also a related discussion in [GR11].*

**Privacy of $v_i$.** The mechanisms presented in [GR11] for selling private information do not protect the privacy of $v_i$ nor they account for the information (dis)utility generated by the leakage of $v_i$. It is further shown that without setting a bound on $v_i$ it is impossible to construct mechanisms that compensate agents for their loss in privacy and achieve reasonable accuracy (in the sense that enough agents sell their information).

Our mechanisms provide an intermediate solution. First, we provide $\varepsilon$-differential privacy to *all* agents, where the guarantee is with respect to their combined type, i.e., $(t_i, v_i)$, and where $\varepsilon$ decreases with the number of agents $n$. This means that privacy improves as $n$ grows.

Furthermore our constructions guarantee that truthfulness is dominant – *taking information utility about the combined type $(t_i, v_i)$ into account* – for all agents for which $v_i \leq v_{max}$, where under a very mild assumption on the distribution of $v_i$ the bound $v_{max}$ grows with $n$ and the fraction of agents for which $v_i > v_{max}$ decreases with $n$.

# 4 The Model

**The Mechanism.** Let $S$ be a finite set of alternatives (a.k.a. social alternatives), let $T$ be a finite type set and consider a set of $n$ agents. We consider direct revelation mechanisms that given the declaration of agents about their types selects a social alternative $s \in S$ and makes $s$ public. To isolate loss of privacy due to publication of $s$ from other potential sources of leakage, we will assume that all other information (including, e.g., the agents' declared types and individual monetary transfers) is completely hidden using cryptographic or other techniques. For example, in an auction of an unlimited supply good (such as is elaborated in Example 5.5 appearing later in the paper), the mechanism solicits agents' bids, and then sets a price $s$. Whether an agent pays $s$ and receives the digital good can be kept hidden using cryptography.

**The Objective Function.** The goal of the designer is defined via a real, non-negative objective function over the true types of the agents, $f(t, s)$ that needs to be optimized (by choosing $s$).

$$f : T^n \times S \to \mathbb{R}^{\geq 0}.$$

Following [DMNS06, MT07] we define the sensitivity of $f$ to be

$$\Delta f = \max \left| f(\hat{t}, s) - f(\hat{t}', s) \right|$$

where the maximum is taken over all neighboring $\hat{t}, \hat{t}' \in T^n$ and $s \in S$. We further assume that for all $s$ the minimum value of $f(t, s)$ is 0 and then, given that sensitivity is $\Delta f$ we get that for all $t \in T^n$, $s \in S$,

$$f(t, s) \in [0, n\Delta f].$$

**Privacy-Aware Agents.** We extend the traditional setting of selfish agents to include agents who care not only about their utility $u_i^{\text{out}}$ from the outcome $s$ of the mechanism, but also about the (negative) information utility $u_i^{\text{inf}}$ incurred from the leakage of information about their private type through the public output $s$.

For simplicity, we consider a setting where the overall utility of an agent is the sum of the two:[2]

$$u_i = u_i^{\text{out}} - u_i^{\text{inf}}.$$

An agent's type $\tau_i$ is modeled by a pair $\tau_i = (t_i, v_i) \in T \times \mathbb{R}^{\geq 0}$, where $T$ is the "traditional" game type and $v_i$ is the privacy valuation of agent $i$. We emphasize that agents care about the privacy of the whole pair and the information utility corresponds to the loss in privacy of both $t_i$ and $v_i$ (hence, one cannot simply publish $v_i$). The vectors $t = (t_1, \ldots, t_n)$ and $v = (v_1, \ldots, v_n)$ denote the types of all agents. Trying to maximize her utility, agent $i$ may hence act strategically and declare $\tau_i' = \sigma_i(\tau_i) = (t_i', v_i')$ to $M$ instead of $\tau_i$.

The "traditional" game utility of agent $i$ is defined as $u_i^{\text{out}} : T \times S \to [-1, 1]$. Following our discussion above, we define $u_i^{\text{inf}} : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$, and the only assumption we make is that

$$u_i^{\text{inf}}(v_i) \leq v_i \cdot \varepsilon$$

for $\varepsilon$ being the parameter of the differentially private mechanism executed (i.e., $e^{\varepsilon} = \max\left(M(t)(S)/M(t')(S)\right)$ where the maximum is taken over all neighboring $t, t' \in T^n$ and $S' \subseteq S$). We emphasize that unlike $u_i^{\text{out}}$ that only depends on the outcome of the mechanism, $u_i^{\text{inf}}$ depends on the mechanism itself.

In our analysis we identify a subset of agents that we call *participating* for whom truthtelling is strictly dominant. A mechanism approximately implements $f$ if assuming that participating agents act truthfully (and other agents act arbitrarily) it outputs $s$ that approximately optimizes $f$.

## 4.1 Warmup: A Privacy-Aware Poll

The following simple electronic poll will serve to illustrate some of our ideas:

**Example 4.1** (An Electronic Poll)**.** *An electronic publisher wishes to determine which of its $m \geq 2$ electronic magazines is more popular. Every agent is asked to specify her favorite magazine, i.e., $t_i \in [m]$, and will receive in exchange an electronic copy of it. For simplicity, we assume that agents' utility depends only on the magazine she receives and not on the poll outcome.*

*Following our convention, we assume ideal cryptography here, that is, no information beyond the outcome of the poll is leaked. In particular, every agent receives the electronic magazine without anybody (including the publisher) knowing which magazine has been transferred. Agents, however, are privacy-aware, and hence take into account that the outcome of the poll itself reveals information about their preferences.*

Denote by $t'$ the vector of agents' declarations. For $s \in [m]$ let $f(s, t') = |\{i : t_i' = s\}|$ and note that $\Delta f = 1$. Consider the exponential mechanism $M = M_f^{\frac{\varepsilon}{2}}$ as in Definition 2.3. I.e.,

$$\Pr[M(t') = j] = \frac{e^{\varepsilon n_j'/2}}{\sum_{\ell=1}^{m} e^{\varepsilon n_\ell'/2}},$$

where $n_j' = f(j, t')$ is the number of agents who declared they rank magazine $j$ first. By Theorem 2.4, $M$ preserves $\varepsilon$-differential privacy.

Note that if $n_j' \geq n_\ell' + k$ then

$$\Pr[M(t') = \ell] \leq \frac{e^{\varepsilon n_\ell'/2}}{e^{\varepsilon n_j'/2}} \leq \frac{e^{\varepsilon n_\ell'/2}}{e^{\varepsilon (n_\ell' + k)/2}} = e^{-\varepsilon k/2}.$$

[2] Admittedly, this separation of the utility function is sometimes artificial. However, we find it conceptually helpful.

Hence,

$$\Pr[M(t') \text{ outputs } \ell \text{ such that } n'_\ell < \max_j n'_j - k] \leq (m-1)e^{-\varepsilon k/2}.$$

The utility of agent $i$ is $u_i^{\text{out}} - u_i^{\text{inf}}$ where

- $u_i^{\text{out}}$ is the utility that the agent gains from receiving the magazine she specified she prefers. Note that this utility depends only on the declared type and it is maximized for $t_i$, the true type of the agent; we assume that $u_i^{\text{out}}(t_i) - u_i^{\text{out}}(t'_i) \geq g$ (Alternatively, the publisher does not care if agent $i$ reports $t'_i$ if $u_i^{\text{out}}(t_i) - u_i^{\text{out}}(t'_i) < g$).
- $u_i^{\text{inf}} \leq \varepsilon \cdot v_i$ is the privacy loss from the mechanism.

Note that $\varepsilon < g/v_i$ suffices for making agent $i$ truthful: acting untruthfully agent $i$ will lose at least $g$ in $u_i^{\text{out}}$ and gain no more than $\varepsilon v_i$ in $u_i^{\text{inf}}$. Denote by $n_j$ the number of agents who rank magazine $j$ first (note the difference from $n'_j$ that correspond to declared types). To demonstrate that the mechanism is efficient, we need to make some (hopefully reasonable) assumptions on the distribution of $v_i$. We explore three possibilities:

**Bounded $v_i$.** We begin with a simplified setting where we assume that there exists $v_{max} = O(1)$ such that $\forall i : v_i \leq v_{max}$. In this case it is enough to set $\varepsilon < g/v_{max} = O(1)$ to make truthfulness dominant for *all* agents. Hence, assuming all agents are truthful, we get $n'_j = n_j$ for all $j \in [m]$ and hence the probability that $M(t') = M(t)$ outputs $\ell$ such that $n_\ell < \max_j n_j - k$ is bounded by $(m-1)e^{-\varepsilon k/2}$. Note that in this case the computation output leaks no information about the privacy valuations $v$.

**Bounded $v_i$, Except for a Small Number of Agents.** A more realistic setting allows for a small number of agents with $v_i > v_{max}$. We change the mechanism $M$ to also consider the reported $v'_i$ so that inputs from agents with $v'_i \geq v_{max}$ are ignored. Regardless of what agents with $v_i > v_{max}$ report, we call them *non-participating*.

As before, by setting $\varepsilon < g/v_{max}$ we make truthfulness dominant for all agents with $v_i \leq v_{max}$. We can hence guarantee a non-trivial accuracy. Let $n_{np}$ be the number of non-participating agents. As these agents may choose not to report truthfully, in the worst case they can deflate the count of a popular magazine and inflate the count of an unpopular magazine, making it look more popular than it really is. Taking this into account, we get that

$$\Pr[M(t') \text{ outputs } \ell \text{ such that } n_\ell < \max_j n_j - k - 2n_{np}] \leq (m-1)e^{-\varepsilon(k+2n_{np})/2}.$$

Note that we lose truthfulness for non-participating agents. We do, however, guarantee $\varepsilon$-differential privacy for these agents.

**Large Populations.** Assume we do not care if the mechanism does not output the most popular choice if it does not have significant advantage over the other, e.g., when $k + n_{np} = O(n^\alpha)$ for some $0 < \alpha < 1$. This allows us to set $\varepsilon(n) = n^{-\alpha}$ and hence truthfulness is dominant for agents with $v_i \leq g/\varepsilon = v_{max}(n) \in O(n^\alpha)$. Note that $v_{max}$ grows with $n$, hence we expect the fraction of non-participants $n_{np}/n$ to diminish with $n$. If $n$ is large enough so that the fraction of agents for which $v_i > v_{max}(n)$ is at most $1/n^{1-\alpha}$ then we get the desired accuracy.

As before, we lose truthfulness for non-participating agents, and only guarantee $\varepsilon(n)$-differential privacy for the non-participating agents. Note, however, that the fraction of non-participating agents diminishes with $n$, and, furthermore, their privacy guarantee improves with $n$ (i.e., $\varepsilon(n)$ decreases).

### 4.2 Admissible Privacy Valuations

In the rest of the paper we only focus on large populations. We will design our mechanisms for "nicely-behaving" populations:

**Definition 4.2** (Admissible Valuations)**.** *A population of $n$ agents is said to have* $(\alpha, \beta)$-admissible valuations *if*

$$\frac{|\{i : v_i > n^\alpha\}|}{n} \leq n^{-\beta}.$$

To partly justify our focus on admissible valuations, consider the case where $v_i$ are chosen, i.i.d., from some underlying distribution $\mathcal{D}$ over $\mathbb{R}^{\geq 0}$.

**Definition 4.3** (Admissible Valuation Distribution)**.** *A valuation distribution $\mathcal{D}$ is called* $(\alpha, \beta)$-admissible *if*

$$\Pr_{v \sim \mathcal{D}}[v > n^\alpha] = O(n^{-\beta}).$$

Note that if $\mathcal{D}$ has finite expectation, then (using Markov's inequality) $\Pr[v > n^\alpha] \leq \mathbf{E}[v]/n^\alpha = O(n^{-\alpha})$, and hence $\mathcal{D}$ is $(\alpha, \beta)$-admissible for all $\beta \leq \alpha$. If $\mathcal{D}$ has finite variance then (using Chebyshev's inequality) $\Pr[v > n^\alpha] \leq \mathbf{Var}[v]/(n^\alpha - \mathbf{E}[v])^2 = O(n^{-2\alpha})$, and hence $\mathcal{D}$ is $(\alpha, \beta)$-admissible for all $\beta \leq 2\alpha$. More generally, consider the following simple generalization of Chebyshev's inequality to even $p$-th moment:

$$\Pr[|X - \mathbf{E}[X]| > t] = \Pr[(X - \mathbf{E}[X])^p > t^p] \leq \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^p\right]}{t^p}.$$

Using this inequality in the argument above we get that if $\mathcal{D}$ has finite even $p$-th moment then $\mathcal{D}$ is $(\alpha, \beta)$-admissible for all $\beta \leq p\alpha$. We conclude that if $\mathcal{D}$ has finite $p$-th moment then $\mathcal{D}$ is $(\alpha, 1-\alpha)$-admissible for $\alpha \geq 1/(p+1)$. In particular, if $\mathcal{D}$ has finite moments of all orders then $\mathcal{D}$ is $(\alpha, 1-\alpha)$-admissible for all $\alpha \in (0, 1)$. These simple observations suggest that $(\alpha, 1-\alpha)$-admissibility is a relatively mild assumption that would typically hold in large populations even for small values of $\alpha$.

It is reasonable to consider even a stronger notion of admissibility:

**Definition 4.4** (Strongly Admissible Valuation Distribution)**.** *A valuation distribution $\mathcal{D}$ is called* $\alpha$-strongly admissible *if*

$$\Pr_{v \sim \mathcal{D}}[v > (\log n)^\alpha] = n^{-\omega(1)},$$

*where $n^{-\omega(1)}$ denotes a function that is negligible in $n$.*

For example, the Normal distribution is strongly admissible. In our analysis, however, we only use the more conservative notion of admissibility as in definitions 4.2 and 4.3.

### 4.3 The Privacy-Aware Poll with Admissible Valuations

Returning to our example, let $\alpha$ be the smallest positive value such that the agent population can be assumed to be $(\alpha, 1-\alpha)$-admissible. By setting $v_{max} = n^\alpha$ and $\varepsilon = g/v_{max} = gn^{-\alpha}$ we get that $n_{np} \leq n \cdot n^{-(1-\alpha)} = n^\alpha$. Finally, setting $k = n^\alpha (\log n)^2 \log m/g$ we get the following:

**Claim 4.5.** *The probability that $M(t')$ outputs $\ell$ such that $n_\ell < \max_j n_j - 2k$ is negligible in $n$.*[3]

## 5 A Generic Construction of Privacy-Aware Mechanisms

We now present a generic feasibility result for privacy-aware mechanisms. Our construction is based on the construction of [NST12], where differential privacy is used as a tool for mechanism design. The hope is that the existence of this generic construction, a relatively simple modification of [NST12], is a signal that our model of privacy-aware mechanisms allows constructing mechanisms for many other tasks.

[3] A function $f(n) : \mathbb{N} \to \mathbb{R}$ is negligible in $n$ if for every positive polynomial $p(n)$ there exists $n_0 \in \mathbb{N}$ such that $f(n) \leq 1/p(n)$ for all $n \geq n_0$. In words, a negligible $f(n)$ is asymptotically smaller than any inverse polynomial.

**Reactions.** We first change our model to incorporate the notion of *reactions* introduced in [NST12].[4] Traditionally, an agent's utility is a function of her private type and the social alternative, and the issue of how agents exploit the social choice is not treated explicitly. In [NST12] this choice was made explicit such that after a social choice is made agents need to take an action (denoted *reaction*) to exploit the social alternative and determine their utility. In [NST12] (and likewise in this work) allowing the mechanism to sometimes restrict the reactions of agents serves as a deterrent against non-truthful agents.

Let $R$ be a finite set of reactions. We modify the definition of the utility from the outcome of the mechanism to

$$u_i^{\text{out}} : T \times S \times R \to [-1, 1].$$

Given $t_i, s$ define

$$r_i(t_i, s) = \text{argmax}_{r \in R}(u_i^{\text{out}}(t_i, s, r))$$

to be the *optimal reaction* for agent $i$ on outcome $s$.

To illustrate the concept of reactions, consider a mechanism for setting a price for an unlimited supply good (such as in Example 5.5 appearing below). Once the mechanism chooses a price $s$ the possible reactions are **buy** (i.e., pay $s$ and get the good) and **not buy** (i.e., do not pay $s$ and do not get the good), and reactions are kept hidden by assuming payment and reception of the digital good using perfect cryptography. In this example agents reactions may be restricted to **buy** whenever they bid at least the selected price $s$, and **not buy** otherwise.

**Utility Gap.** We assume the existence of a positive *gap* $g$ such that for all $t_i \neq t_i'$ there exist $s$ for which the optimal reactions are distinct, and, furthermore,

$$u_i^{\text{out}}(t_i, s, r_i(t_i, s)) \geq u_i^{\text{out}}(t_i, s, r_i(t_i', s)) + g.$$

In many settings, a gap $g$ can be created by considering a discrete set of social choices. As in our polling example, an alternative interpretation of the gap $g$ may be that the mechanism designer does not care if agent $i$ reports $t_i'$ if $u_i^{\text{out}}(t_i, s, r_i(t_i, s)) < u_i^{\text{out}}(t_i, s, r_i(t_i', s)) + g$.

## 5.1 The Construction

Given a finite type set $T$, a finite set $S$ of alternatives and an objective function $f : T^n \times S \to \mathbb{R}$ with sensitivity $\Delta f$, we construct a mechanism for approximately implementing $f$.

Let $n$ be the number of agents, and let $\alpha$ be the smallest positive value such that the agent population can be assumed to be $(\alpha, 1-\alpha)$-admissible. Let $v_{max} = n^\alpha$. The participating agents will be those with privacy valuations lower than $v_{max}$. Choose $t_\perp \in T$ to be an arbitrary element of $T$. Non-participating agents will be asked to declare $t_\perp$.

Let $\delta \in [0, 1], \varepsilon > 0$ be parameters to be set later. Agents are asked to declare $t_i$ if $v_i \leq v_{max}$ and $t_\perp$ otherwise. Let $t_i'$ be the declaration of agent $i$. On input $t' = t_1', \ldots, t_n'$ the mechanism runs as described in Algorithm 1.

We begin by analyzing for which agents truthtelling is a dominant strategy:

**Claim 5.1.** *If* $(v_{max} + 4)\varepsilon \leq \delta \frac{g}{|S|}$ *then truthtelling is dominant for all agents with* $v_i \leq v_{max}$.

*Proof.* We first analyze the effect of misreporting in $M_1$ and $M_2$:

[4] While the standard game-theoretic modeling does not explicitly include reactions, in many settings their introduction is natural. We refer the reader to [NST12] for further discussion of this change in the standard model.

[5] It suffices to restrict reactions only when $M_2$ is activated.

**ALGORITHM 1:** The generic mechanism $M$.

**Input**: A vector of types $t' \in T^n$.
**Output**: A social choice $s \in S$.
$M$ executes $M_1$ with probability $1-\delta$ and $M_2$ otherwise, where $M_1, M_2$ are as follows:

**Mechanism $M_1$:** For all $s \in S$ and $t' \in T^n$, choose $s \in S$ according to the exponential mechanism $M_f^{\frac{\varepsilon}{2\Delta f}}(t')$.

**Mechanism $M_2$:** Choose $s \in S$ uniformly at random.

The mechanism $M$ also restrict all agents to their optimal reactions according to their declarations, i.e., $r_i(t'_i, s)$.[5]

**Misreporting in $M_1$:** As $u_i^{\text{out}}(t_i, s, r) \in [-1, 1]$ we can use the simple corollary following Lemma 2.2 and get that for all possible declarations of the other agents $t'_{-i}$ and all $t'_i$:

$$\begin{aligned}\mathbf{E}_{s\sim M_1(t'_{-i},t'_i)}[u_i^{\text{out}}(t_i, s, r_i(t'_i, s))] - \mathbf{E}_{s\sim M_1(t'_{-i},t_i)}[u_i^{\text{out}}(t_i, s, r_i(t_i, s))] &\leq \\ \mathbf{E}_{s\sim M_1(t'_{-i},t'_i)}[u_i^{\text{out}}(t_i, s, r_i(t'_i, s))] - \mathbf{E}_{s\sim M_1(t'_{-i},t_i)}[u_i^{\text{out}}(t_i, s, r_i(t'_i, s))] &< 4\varepsilon,\end{aligned}$$

where the first inequality follows from $u_i^{\text{out}}(t_i, s, r_i(t'_i, s)) \leq u_i^{\text{out}}(t_i, s, r_i(t_i, s))$. In words, misreporting can gain at most $4\varepsilon$ in the expected $u_i^{\text{out}}$.[6] Noting that misreporting can gain agent $i$ at most $v_i \cdot \varepsilon$ in $u_i^{\text{inf}}$, we get that the total gain in utility due to misreporting by agents with $v_i \leq v_{max}$ is $(v_{max}+4)\varepsilon$.

**Misreporting in $M_2$:** If $t'_i \neq t_i$ then with probability at least $\frac{1}{|S|}$ we get that $r_i(t'_i, s) \neq r_i(t_i, s)$. Since the mechanism restricts agent $i$'s reaction to $r_i(t'_i, s)$ we get that

$$\mathbf{E}_{s\sim M_2(t'_{-i},t_i)}[u_i^{\text{out}}(t_i, s, r_i(t_i, s))] - \mathbf{E}_{s\sim M_2(t'_{-i},t'_i)}[u_i^{\text{out}}(t_i, s, r_i(t'_i, s))] \geq \frac{g}{|S|},$$

where $g$ is the minimal utility gap due to not acting according to the optimal reaction. Note that, as $M_2$ ignores its input, misreporting does not yield a change in $u_i^{\text{inf}}$. We get that the total loss in utility in $M_2$ due to misreporting is at least $g/|S|$.

We get that if $(v_{max}+4)\varepsilon \leq \delta\frac{g}{|S|}$ then overall gain in utility due to misreporting is negative for all agents with $v_i \leq v_{max}$, hence truthtelling is dominant for these agents. □

Let $\text{opt}(t) = \max_{s\in S} f(t, s)$ be the optimal value for $f$. We next show that our mechanism approximately recovers $\text{opt}(t)$.

**Claim 5.2.** *If $(v_{max}+4)\varepsilon \leq \delta\frac{g}{|S|}$ then*

$$\mathbf{E}_{s\sim M(t')}[f(t, s)] \geq \text{opt}(t) - \Delta f \cdot (\delta n + 2n^{\alpha} + 2\ln(n|S|)/\varepsilon)\,.$$

*Proof.* Define $\text{opt}' = \max_s f(t', s)$. Denote by $\bar{t}_k$ the vector constructed from the $k$ first entries of $t$ and the $n-k$ last entries of $t'$. For all $s$ we have that

$$f(t', s) = f(\bar{t}_0, s) = \sum_{k=0}^{n-1}(f(\bar{t}_k, s) - f(\bar{t}_{k+1}, s)) + f(t, s).$$

Note that by Claim 5.1 $t'_i \neq t_i$ for at most $n^{\alpha}$ entries, and hence $f(\bar{t}_k, s) - f(\bar{t}_{k+1}, s) \neq 0$ for at most $n^{\alpha}$ values of $k$, in which case it is upper bounded by $\Delta f$. We get hence that $\text{opt}' \geq \text{opt}(t) - n^{\alpha}\Delta f$.

We get that $M_1(t')$ outputs $s'$ such that $f(t', s') < \text{opt}' - 2\Delta f \ln(n|S|)/\varepsilon$ with probability

$$\frac{\exp(\varepsilon f(t', s')/2\Delta f)}{\sum_{s\in S}\exp(\varepsilon f(t', s)/2\Delta f)} \leq \frac{\exp(\varepsilon(\text{opt}' - 2\Delta f\ln(n|S|)/\varepsilon)/2\Delta f)}{\exp(\varepsilon\text{opt}'/2\Delta f)} = \frac{1}{n|S|}.$$

[6] Similarly, even if reactions are not restricted when $M_1$ is activated we get that: $\mathbf{E}_{s\sim M_1(t'_{-i},t'_i)}[u_i^{\text{out}}(t_i, s, r_i(t_i, s))] - \mathbf{E}_{s\sim M_1(t'_{-i},t_i)}[u_i^{\text{out}}(t_i, s, r_i(t_i, s))] < 4\varepsilon$. We only need reactions to be restricted when $M_2$ is activated.

Using the union bound (over elements of $S$), and the fact that $\text{opt}' \leq n\Delta f$, we get a lower bound on the expected revenue of $M_1$ as follows:

$$\begin{aligned}\mathbf{E}_{s\sim M_1(t')}[f(t,s)] &\geq (\text{opt}' - 2\Delta f \ln(n|S|)/\varepsilon)\left(1 - |S|\frac{1}{n|S|}\right)\\ &\geq \text{opt}' - 2\Delta f \ln(n|S|)/\varepsilon - \Delta f\\ &\geq \text{opt}(t) - 2n^{\alpha}\Delta f - 2\Delta f \ln(n|S|)/\varepsilon.\end{aligned}$$

We conclude that

$$\begin{aligned}\mathbf{E}_{s\sim M(t')}[f(t,s)] &\geq (1-\delta)\mathbf{E}_{s\sim M_1(t')}[f(t,s)]\\ &\geq (1-\delta)\left(\text{opt}(t) - 2n^{\alpha}\Delta f - 2\Delta f \ln(n|S|)/\varepsilon\right)\\ &\geq \text{opt}(t) - \delta n\Delta f - 2n^{\alpha}\Delta f - 2\Delta f \ln(n|S|)/\varepsilon.\end{aligned}$$

□

Setting $\varepsilon = n^{-(1+\alpha)/2}\sqrt{g\ln(n|S|)/|S|}$ and $\delta = 2n^{(\alpha-1)/2}\sqrt{|S|\ln(n|S|)/g}$ we get

**Theorem 5.3.** *Let $n$ be the number of agents, $T$ be a finite type set and $S$ a finite set of alternatives. Let $f : T^n \times S \to \mathbb{R}$ be an objective function with sensitivity $\Delta f$ and $M$ be the mechanism described in Algorithm 1.*

*If $\alpha$ is such that the agent population can be assumed to be $(\alpha, 1-\alpha)$-admissible, then $M$ recovers* $\text{opt}(t)$ *to within additive difference of $O\left(\Delta f n^{(1+\alpha)/2}\sqrt{|S|\ln(n|S|)/g}\right)$.*

As discussed above, natural distributions of the privacy valuations will be $(\alpha, 1-\alpha)$-admissible even for very small values of $\alpha$, and therefore the dominating term in the expression in Theorem 5.3 can be made arbitrarily close to $\tilde{O}(\sqrt{n})$.

**Note 5.4.** *The proof of Theorem 5.3 only considers pure (i.e., deterministic) strategies. Extending the theorem to mixed strategies requires a proper relaxation of the solution concept, namely, to implementation in* undominated strategies.

## 5.2 Example: Privacy-Aware Selling of Digital Goods

We describe now an example of a natural privacy-aware mechanism that naturally falls within our framework.

**Example 5.5** (Pricing a Digital Good)**.** *An auctioneer selling a digital good wishes to design a single price mechanism that would (approximately) optimize her revenue. Every party has a valuation $t_i \in Q = \{0, \frac{1}{q}, \frac{2}{q}, \ldots, 1\}$ for the good (for some constant $q$), and a privacy preference $v_i$. Agents are asked to declare $t_i$ to the mechanism, which chooses a price $p$ for the good. Denote by $t'_i$ the actual declaration of agent $i$. If $t'_i \geq p$ then agent $i$ receives the good and pays $p$, otherwise, agent $i$ learns $p$ but does not pay nor receive the good. Agents prefer receiving the good to not receiving it.*

- *The utility $u^{\text{out}}$ is the 'traditional' utility, i.e., zero if agent $i$ does not receive the good, and $t_i - p + \frac{1}{2q}$ otherwise, where the additive $\frac{1}{2q}$ is used for modeling preference to receive the good.*
- *For $u^{\text{inf}}$, we assume that whether agent $i$ received the good and paid for it can be kept completely hidden from all other parties (this can be implemented using cryptographic techniques). Hence, only leakage due to making $p$ public affects $u^{\text{inf}}$.*

Consider now the auctioneer from Example 5.5, and assume that the valuations $t_i$ are taken from $T = \{0, \frac{1}{q}, \frac{2}{q}, \ldots, 1\}$ and similarly, the price $p \in S = \{\frac{1}{q}, \frac{2}{q}, \ldots, 1\}$ for some integer constant $q > 1$. Let $\alpha$ be the smallest value such that the agent population can be assumed to be $(\alpha, 1-\alpha)$-admissible.

Defining the reactions to be $\{\textbf{buy}, \textbf{not buy}\}$ and optimal reactions $r_i(t,p) = \textbf{buy}$ if $t \geq p$ and **not buy** otherwise we get that the gap $g$ is $1/2q$.

Suppose the designer goal is to recover the optimal revenue, i.e., $\max_{p \in S} f(t,p)$ where $f(t,p) = p \cdot |\{i : t_i \geq p\}|$ and note that $\Delta f = 1$.

Using Theorem 5.3 we get a privacy-aware mechanism that recovers the optimal revenue to within additive difference of $O\left(\Delta f n^{(1+\alpha)/2}\sqrt{|S|\ln(n|S|)/g}\right) = O\left(n^{(1+\alpha)/2} q \sqrt{\ln(nq)}\right)$.

Note that the accuracy of this privacy aware mechanism is only (essentially) a factor $\tilde{O}(n^{\frac{\alpha}{2}})$ away from the similar (non-privacy aware) mechanism from [NST12].

# References


[AR11] Itai Ashlagi and Alvin Roth. Individual rationality and participation in large scale, multi-hospital kidney exchange. In Shoham et al. [SCR11], pages 321–322.

[CCK+11] Yiling Chen, Stephen Chong, Ian A. Kash, Tal Moran, and Salil P. Vadhan. Truthful mechanisms for agents that value privacy. *CoRR*, abs/1111.5472, 2011.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, John Wiley and Sons, Inc., 1991.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, pages 51–60. IEEE Computer Society, 2010.

[GR11] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In Shoham et al. [SCR11], pages 199–208.

[MNT09] Peter Bro Miltersen, Jesper Buus Nielsen, and Nikos Triandopoulos. Privacy-enhancing auctions using rational cryptography. In Shai Halevi, editor, *CRYPTO*, volume 5677 of *Lecture Notes in Computer Science*, pages 541–558. Springer, 2009.

[MPS03] Robert McGrew, Ryan Porter, and Yoav Shoham. Towards a general theory of non-cooperative computation. In Joseph Y. Halpern and Moshe Tennenholtz, editors, *TARK*, pages 59–71. ACM, 2003.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103. IEEE Computer Society, 2007.

[NPS99] Moni Naor, Benny Pinkas, and Reuban Sumner. Privacy preserving auctions and mechanism design. In *ACM Conference on Electronic Commerce*, pages 129–139, 1999.

[NST12] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. Approximately optimal mechanism design via differential privacy. *Innovations of Theoretical Computer Science (ITCS)*, 2012. Electronic version available in *CoRR*, abs/1004.2888, 2010.

[SCR11] Yoav Shoham, Yan Chen, and Tim Roughgarden, editors. *Proceedings 12th ACM Conference on Electronic Commerce (EC-2011), San Jose, CA, USA, June 5-9, 2011*. ACM, 2011.

[ST05] Yoav Shoham and Moshe Tennenholtz. Non-cooperative computation: Boolean functions with correctness and exclusivity. *Theor. Comput. Sci.*, 343(1-2):97–113, 2005.

[WZ04] Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

[Xia11] David Xiao. Is privacy compatible with truthfulness? Cryptology ePrint Archive, Report 2011/005, 2011. `http://eprint.iacr.org/`.