Is Privacy Compatible with Truthfulness?

David Xiao
LIAFA
CNRS, Université Paris 7
Paris, France
dxiao@liafa.univ-paris-diderot.fr

November 23, 2012

Abstract

In the area of privacy-preserving data mining, a differentially private mechanism intuitively encourages people to share their data because they are at little risk of revealing their own information. However, we argue that this interpretation is incomplete because external incentives are necessary for people to participate in databases, and so data release mechanisms should not only be differentially private but also compatible with incentives, otherwise the data collected may be false. We apply the notion of *truthfulness* from game theory to this problem. In certain settings, it turns out that existing differentially private mechanisms do not encourage participants to report their information truthfully.

On the positive side, we exhibit a transformation that takes truthful mechanisms and transforms them into differentially private mechanisms that remain truthful. Our transformation applies to games where the type space is small and the goal is to optimize an insensitive quantity such as social welfare. Our transformation incurs only a small additive loss in optimality, and it is computationally efficient. Combined with the VCG mechanism, our transformation implies that there exists a differentially private, truthful, and approximately efficient mechanism for any social welfare game with small type space.

We also study a model where an explicit numerical cost is assigned to the information leaked by a mechanism. We show that in this case, even differential privacy may not be strong enough of a notion to motivate people to participate truthfully. We show that mechanisms that release a perturbed histogram of the database may reveal too much information. We also show that, in general, any mechanism that outputs a synopsis that resembles the original database (such as the mechanism of Blum et al. (STOC '08)) may reveal too much information.

1 Introduction

As our world becomes ever more digitized and connected, concerns about the privacy of our personal information have become increasingly pressing. With various organizations collecting, accessing, and storing individuals' data, our desire to fully take advantage of the data has come into stark tension with the equally important desire to keep information about various aspects of our lives private, including for example our medical histories and our demographic information.

Understanding and resolving this tension has been a long-standing goal in the statistics and data analysis literature [5, 16]. More recently, the theoretical computer science community has provided definitions and a rigorous treatment of this question [7, 8, 11, 3, 12], culminating in the definition and study of differential privacy [11, 8]. See [9] for a more complete overview of the area and further references.

Differential privacy guarantees the following: a (randomized) data release mechanism is ε-differentially private if, for any input database, any participant i in the database, and any possible output of the release mechanism s, the presence or absence of participant i causes at most a multiplicative e^{ε} change in the probability of the mechanism outputting s. The question is whether one can achieve privacy while at the same time preserving some notion of utility, since one could always build a trivially private (but not very useful) mechanism that outputs a constant value without looking at the input data. The advantage of working with the rigorous definition provided by differential privacy is that one can now show that certain tasks can be done with formal guarantees of utility and privacy [2, 3, 24, 27, 15, 29].

For a problem where one can achieve both meaningful utility and differential privacy, the differential privacy guarantee is interpreted as follows: since the output distribution of the mechanism is barely affected by the presence or absence of participant i in the above very precise and very strong sense, therefore the release mechanism leaks very little private information about participant i, and participant i should feel comfortable entering his data in the database.

Differential privacy and incentives In this paper we investigate this interpretation. The first question we pose is why do individuals want to participate in a database at all? If individuals are indeed concerned about the privacy of their information, then they must receive some incentive to participate in order to counteract the preference for guarding their information private.

It seems hard to imagine individuals submitting private data to an organization that gives them no incentive whatsoever (interpreting "incentive" broadly). For example, individuals do not willingly submit their information to marketing companies selling something they have no interest in (unless they are incentivized by money or the chance of winning a prize). On the other hand, when something can be gained, such as the pleasure or connection to friends offered by Facebook or by using email hosted on Google, individuals do readily reveal private information. Approaching incentives from the point of view of game theory, we therefore posit that for any database where the information being collected is sensitive, there must be an associated game whose outcome incentivizes the individuals to participate in the database.

We will study the most general setting of incentives in the following two senses: (1) the incentives of the individual may be incompatible with each other and with the incentives of the database curator, and (2) the individuals may be able to lie about their data if it is to their advantage. See Section 2.4.1 for a discussion of work that uses different assumptions.

2 Our results

The main contribution of this paper is conceptual: we advocate combining the study of differential privacy and game-theoretic incentives into the following natural and coherent model. We also prove some general positive results in this model showing that one can achieve differential privacy in an incentive-compatible way, and we give negative results showing that additional care must be taken when assigning explicit costs to privacy.

2.1 Our model

The differential privacy setting There is a database curator who solicits information from individuals and puts them into a database. Each row of the database contains the information of one individual. The curator then applies some *release mechanism* to the database in order to produce an output, which is then published. We assume that there is some *quality metric* that determines how good the output of the release mechanism is. The database curator wants to use a mechanism that produces high quality outputs, but the individuals wish for the output to be differentially private, and these two may be in tension.[1]

The mechanism design setting We next give an informal introduction to the game-theoretic setting of mechanism design, and we then relate it to the problem of differential privacy (more precise definitions appear in Section 3). In order to make things slightly more concrete, for each notion that we introduce, we state what it would be in the case of an auction for a single item.

In mechanism design, there are n players, each of whom has some private information called his *type* (*e.g.* in a single-item auction, each player's type would be how much he values the item), and there is a space of possible outcomes (*e.g.* a choice of winner and a price). Each of the n players has a *individual utility function* that determines the amount of utility a player gets from each possible outcome (*e.g.* the difference between the price and his valuation if the player wins the auction, or 0 if he loses). The goal of each individual in participating in the game is to optimize his individual utility. There is also a "global utility" function that

[1] As is typical in the differential privacy literature, we treat the database curator as a trusted third party, who is allowed to see the individuals' private information.

determines the overall quality of the outcome (*e.g.* the social welfare, which is just the sum of the individual utilities of all the players).

A mechanism is a (possibly randomized) function mapping each possible setting of the players' types to an outcome. The main problem studied in mechanism design is to build mechanisms that satisfy the following two properties:

Efficiency For any choice of the players' types, the expected global utility of the output of the mechanism should be close to the optimal global utility.[2]

Truthfulness For any choice of the players' types, no single player, say player i, can report a false type so that (if all other players remain honest) the mechanism gives an outcome with greater expected utility to player i than if player i had reported honestly.

Since individual utilities may not be aligned with the global utility, it turns out that implementing even very simple functionalities satisfying the two criteria above simultaneously is non-trivial. For example, the study of the single-item auction led to the Vickrey 2nd-price auction mechanism.

The combined model We can now see how the differential privacy setting can be comfortably re-stated using game-theoretic terminology. An individual is a player, and each individual's private information is his type. The database is just the concatenation of all players' types. The database release mechanism is just a mechanism in the game-theoretic sense. The quality metric is the global utility of the game, and therefore "accurate" release mechanisms correspond to "efficient" game-theoretic mechanisms. In addition to the above, the *individual* utility functions now may also affect players' behaviors.

Truthfulness is essential not only because it is a standard game-theory notion, but also because privacy does not make sense without truthfulness. If the data being collected were misreported, it would be of little use to the database curator. Moreover, protecting the privacy of false data seems superfluous.

Also, recall that in differential privacy, the concern is that if there is too little privacy then individuals may choose not to participate in a database. In the game theoretic setting, truthfulness and participation are closely related: one can always extend a game to have a "$\perp$" type that represents non-participation, and have a player deviate to the "$\perp$" type to represent non-participation. Conversely, a player who reports a type that is independent of his true type is in some sense choosing not to participate. In the rest of this paper, we will focus on truthfulness, with the understanding that non-participation is a specific kind of deviation from truthfulness.

Roadmap We will see that mechanisms that seem satisfactory when privacy is studied in isolation are not necessarily satisfactory when truthfulness is also explicitly considered.

We will look at the relationship between differential privacy and truthfulness in two frameworks. The first is to construct mechanisms that are truthful, efficient, and whose output satisfies differential privacy. In this model, privacy is not given an explicit numerical value, but is simply an additional property of the mechanism. We call such mechanisms PTE, and we show the first PTE mechanisms by exhibiting a transformation from truthful and efficient mechanisms into PTE mechanisms for games where the type space is small and where the global utility is insensitive to any particular player's type and depends only on the number of players who have each type.

The second framework is where the value of privacy is explicitly quantified. We will assign a non-zero cost to the amount of information leaked by a mechanism, and we ask when does the cost of revealing information overwhelm the utility a player gets from the game. We will show that when privacy has a non-zero cost, even differentially private mechanisms may not motivate individuals to reveal their true information.

While the first framework may seem overly simple by not quantifying privacy loss, the techniques developed there have subsequently been shown to apply to the second framework as well (see Section 2.4.2).

2.2 Private, truthful, and efficient (PTE) mechanisms

The Exponential Mechanism is not necessarily truthful Our first result revisits the question of whether or not the Exponential Mechanism is truthful. Nissim et al. [31] already exhibited a counter-

[2] Throughout this paper, *efficiency* refers to the game-theoretic notion. Computational efficiency will be explicitly stated as such.

example showing that the Exponential Mechanism is not truthful. However, the counter-example they give is somewhat artificial because it is easy to see that *no* efficient and truthful mechanism exists for their game (even without considering privacy). We believe the following Theorem 2.1 highlights the drawback of the Exponential Mechanism in sharper relief than the counter-example of [31], because the LINE-1-FAC game we consider *does* have a truthful and efficient mechanism. In fact, we will see that there even exists a PTE mechanism for LINE-1-FAC.

We consider the well-studied 1-facility location on a line game, denoted LINE-1-FAC, which is a special case of the k-facility problem [31, 26, 1] and also of the single-peaked preference games [28, 33]. In the LINE-1-FAC game, each player has a point on the interval $[0, 1]$ and the mechanism is supposed to output $s \in [0, 1]$. Each individual wants to minimize their distance to s, while the mechanism wants to minimize the sum of all the individuals' distances to s. This game *does* have a truthful and efficient mechanism without money, namely outputting the left-most median player's point [28]. We prove:

Theorem 2.1. *For LINE-1-FAC, the Exponential Mechanism with any privacy $\varepsilon > 0$ is not truthful.*

Transforming a truthful and efficient mechanism into a private, truthful, and efficient mechanism Our main positive technical result is a transformation for a large class of games that converts truthful and efficient mechanisms into PTE mechanisms.

More precisely, we work with the relaxed notion of (ε, η)-differential privacy [10], which, in addition to the e^{ε} multiplicative difference, also allows an additive error η, usually taken to be negligible, in the difference between the output distributions of two databases. Our transformation takes any truthful and efficient mechanism and transforms it into a (ε, η)-differentially private, truthful, and δ-efficient mechanism, where δ is an additive approximation error that depends on ε, η. Formal definitions are deferred to Section 3.

These are the first PTE mechanisms exhibited for non-trivial games. Furthermore, our transformation preserves other properties such as computational efficiency and moneylessness, and it applies to *all* games with small type space. Therefore, applying our transformation to the VCG mechanism gives a PTE mechanism for *all* social welfare games with small type space.

As throughout this paper, we will assume in the following that the utility functions of each user is bounded in $[-1, 1]$, and that the global utility function is bounded between $[-n, n]$. In particular, this means that it is interesting to obtain mechanisms that are δ-efficiency for $\delta = o(n)$.

The transformation The idea of the transformation is based on the ideas for privately releasing histogram data [11]. Suppose the type space of the game is a finite set of size q. For player inputs $\underline{t} = (t_1, \ldots, t_n)$ where t_i is the type of the i'th player, we will let $\underline{h}$ denote the histogram with q entries, one for each possible type, and where $h_j = |\{i \mid t_i = j\}|$, the number of players who have type j.

Suppose each individual player's utility function is $v(t, s)$ where t is the player's type and s is an outcome of the game; suppose that v lies in the interval $[-1, 1]$. Suppose that the global utility function $w(\underline{t}, s)$ is anonymous (*i.e.* it depends only on the histogram $\underline{h}$ and not on player identities), and is insensitive: namely if $\underline{h}, \underline{h}'$ are close in ℓ_1 distance then for all outcomes s it holds that $|w(\underline{h}, s) - w(\underline{h}', s)|$ is small. For example, the social welfare function, which is just the sum of the individual players' utilities, is an example of such a global utility function. Without loss of generality, we can assume that all mechanisms for games with such global utility functions need only depend on the histogram of player types, rather than looking at the individual players' types. (We show this in Appendix A.)

At a high level, our transformation from truthful and efficient to PTE works by constructing the histogram of inputs, adding independent noise distributed according to the two-sided geometric distribution to each of the entries of the histogram, and then running the original truthful and efficient mechanism on the perturbed histogram. Care must be taken that the noise does not create negative entries into the histogram; simply truncating negative entries to 0 does not give truthful mechanisms. We show a way to achieve this in Section 4. Our procedure gives the following theorem.

Theorem 2.2 (Informal)**.** *Let $\varepsilon, \eta > 0$. Let G be a game with a type space of size q and where the global utility function depends only on the histogram and is insensitive to individual players' types. Let M be a truthful and δ-approximately efficient mechanism. Then there exists a truthful mechanism M' for the game G that is (ε, η)-differentially private and is δ'-approximately efficient for $\delta' \approx \delta$. Furthermore, if M is computationally efficient then so is M', and if M is moneyless then so is M'.*

We note that our transformation is applicable to many settings where small type space is natural: for example, in some auction settings, the bids are discretized so there are only a few possible values of bids; in surveys or questionnaires, the questions are often multiple-choice and therefore take on only a few values. In the following we choose to highlight two particularly interesting applications of our transformation.

Application to LINE-1-FAC We will show that using a simple rounding procedure and then applying Theorem 2.2 to a mechansim for the discretized version of the LINE-1-FAC game, we can give a PTE mechanism for LINE-1-FAC.

Corollary 2.3. *For all $\varepsilon, \eta > 0$, there is a (ε, η)-differentially private, truthful, and $O(n^{1/2} \log(n/\eta)/\varepsilon)$ approximately efficient mechanism for LINE-1-FAC.*

Applying our transformation to this game highlights two features: first, although the Exponential Mechanism is not truthful when applied to this game, our transformation shows there does exist a private, truthful, and efficient mechanism for this game. Second, although our main transformation applies to games with small finite type space, our mechanism for LINE-1-FAC shows that in certain cases one can suitably discretize large or infinite type spaces and then apply our transformation.

PTE mechanism for *all* social welfare games with small type space In fact, because the VCG mechanism gives a general truthful and perfectly efficient mechanism (using money) for all social welfare games, *i.e.* games optimizing the sum of the individuals' utilities, our main theorem implies the following:

Corollary 2.4. *Fix $\varepsilon, \eta > 0$. For any game G with n players and where the type space has size q, and where the goal is to optimize social welfare, there is a truthful mechanism for G that is (ε, η)-differentially private and is $O(q \log(q/\eta)/\varepsilon)$-approximately efficient.*

2.3 The value of privacy

Prior work (including the results outlined in the previous section) studied privacy and utility as orthogonal features. For example, in the above we showed that one could achieve a truthful (with respect to utility functions that do not consider privacy) and efficient mechanism that simultaneously satisfied differential privacy. However, if one really believes that participants value their privacy, then this should explicitly be taken into account in their utility functions. Not only does this allow us to quantify *how much* participants value their privacy, it also allows us to model tradeoffs between utility and privacy. That is, participants may be willing to sacrifice some of the utility they reap from a game by reporting a false type and thereby protecting their privacy. Vice versa, participants may be willing to sacrifice their privacy if by doing so they gain a larger utility from the outcome of the game. We will show that it is possible even for PTE mechanisms to leak too much information to achieve truthfulness in games where there exists a tradeoff between utility and privacy.

2.3.1 Quantifying privacy

The first task is to define a measure of how much information a mechanism leaks. One natural condition that the information cost should satisfy is that it should be 0 if a player reports a type that is independent of his honest type, since this means that the output of the mechanism does not contain information about his type, and so should not compromise his privacy. This criterion implies that information cost cannot *solely* be a function of the player's type and the outcome of the game, because the notion "independent of the honest type" is inherently a statement about how the player behaves on *all* possible values of his type. As a thought experiment, fix any type t, and consider the following two strategies: first is the truthful strategy, and second the strategy that always outputs t regardless of what the player's actual type is. Then, in the case that the player's actual type is t, the two strategies give exactly the same output, but intuitively the truthful strategy might reveal information about the player's type while the constant t strategy does not.

Therefore our measure of information cost cannot be expressed as a modification of a "traditional" utility function that depends solely on the player's type and outcome. Instead, we work with a measure that depends on the player's strategy over all possible types.

We let $\mathsf{IC}_M(\sigma, \underline{t}, i)$ denote the information cost to player i of the strategy σ, given that the (true) input is $\underline{t} = (t_1, \ldots, t_n)$ and that the mechanism is M. A strategy σ is simply a function mapping types to distributions over types, so that if player i has true type t, then to use strategy σ he will sample $t' \leftarrow \sigma(t)$ and declare t' to the mechanism.

In general the information cost can be highly application-specific, and so we prefer to state our results without fixing a specific information cost function. Rather, our results hold for any information cost function that satisfies the following assumptions:

Players can hide their data: for any strategy σ that is independent of its input (*i.e.* $\sigma(t) = \sigma(t')$ for all t, t'), then for all $i, \underline{t}$ it holds that $\mathsf{IC}_M(\sigma, \underline{t}, i) = 0$.

Cost reflects differential privacy: roughly, this condition requires that, if ε is the minimal value such that M is a ε-differentially private mechanism, the information cost cannot be much lower than ε.

We formalize the assumptions above in Section 5. We already justified the first assumption above. Intuitively, the second assumption simply means that the amount of differential privacy (the ε) is "dominated" by the application-specific information cost. In contexts where the information cost did not satisfy the second condition, presumably this means that we have application-specific knowledge about what privacy concerns are relevant, and so it would be more sensible to study mechanisms that target directly the application-specific information cost rather than trying to use differential privacy as a generic tool.

In Appendix C we give some candidate definitions of information cost.

Tradeoffs between the value of the game and the value of privacy Given an information cost function, we would like to explicitly incorporate it into the utility of the player. Let us use the word "game value" or simply "value" to denote the benefit that the player extracts from the outcome of the game without taking into account privacy, and we will let overall utility be the game value minus the information cost, and we will say that such a utility is privacy-sensitive.

2.3.2 Releasing histograms for LINE-1-FAC

Our first negative result draws on the LINE-1-FAC game once again. Corollary 2.3 says that a PTE mechanism for LINE-1-FAC exists, let us denote it by M (see Algorithm 4.10 for the definition of the mechanism). In order to understand the following result, we first give a rough description of M: M first rounds the positions of the players to discrete points, then constructs a histogram of the players' rounded locations, perturbs the histogram, and finally outputs the median point of the perturbed histogram. Suppose now that, instead of outputting the median, the mechanism outputs "more than necessary" and also publishes the entire perturbed histogram of the players' locations as well. Call this mechanism $\hat{M}$. We believe that $\hat{M}$, represents a plausible situation in the real world: an agency gathers data for a single explicit purpose, for example to decide where to build a hospital so as best to serve local residents, but then publishes the entire perturbed histogram data so that it may be of use to other agencies that may want to use it in a different way.

$\hat{M}$ is actually PTE: truthfulness and efficiency remain because the facility location output is the same as M, while privacy also remains because the perturbation of the histogram was designed to render the entire histogram differentially private. However, we show in Theorem 5.6 that if we set the parameters so that the mechanism is $(2\varepsilon, \eta)$-differentially private, then the information cost is greater than ε. On the other hand, we show in Theorem 5.3 that there exist situations where the amount of value that a player loses by ignoring his true type and declaring a fixed type, say 0, is at most $e^{-\Omega(n)}$. This implies the following:

Theorem 2.5 (Informal, see Corollary 5.8)**.** *The PTE mechanism for LINE-1-FAC given by $\hat{M}$ is untruthful when one takes into account the information cost.*

2.3.3 Releasing synopses.

We prove a more general theorem about the information cost of any mechanism that publishes information that can be useful for many different count queries: if Q is the type space of the players and $F : Q \to \{0, 1\}$, then we define the count query $\overline{F}(\underline{t}) = \frac{1}{n} \sum_{i=1}^{n} F(t_i)$. For example, in the case of a census, F might be the set of households with two children, and $\overline{F}(\underline{t})$ would be the fraction of all households that have two children.

We say that a mechanism M is a synopsis generator for a class of predicates $\mathcal{C}$ if, given an input database $\underline{t}$, it outputs a data structure $M(\underline{t})$ from which we can deduce $\overline{F}(\underline{t})$ for many $F \in \mathcal{C}$. It was shown by Blum et al. [3] that it is possible to construct differentially private and accurate synopsis generators. Subsequent work has extended this to achieve better parameters [13, 21].

Here we show that good synopsis generators must leak information:

Theorem 2.6 (Informal, see Theorem 5.11)**.** *If M is a good synopsis generator for a rich class of predicates, and IC_M reflects differential privacy, then for almost all databases, the information cost of revealing a synopsis generated from that database is $\Omega(1)$ with respect to at least one player.*

This shows that synopsis generators such as the one proposed by Blum et al. [3] must inherently reveal a lot of information some player's type, and furthermore, that they must leak information on "most" databases (more precisely, we will show that for any database, there is a nearby one where some player has a high information cost). Interestingly, the proof uses a construction of a combinatorial design to show that, for every $\underline{t}$, there exists an exponentially large family of $\underline{t}'$ that differ little from $\underline{t}$ such that one of them has a player with large information cost.

This theorem leads to the following consequences: consider a mechanism M that publishes an output that (a) incentives players' participation by giving them some value and (b) includes a good synopsis. If the value to each player diminishes by very little (say vanishingly small) when a player mis-reports his type, but the information cost is constant (as is implied by Theorem 2.6), then the mechanism cannot be truthful. Namely, individuals will prefer to lie because by declaring, say, a constant value, they will gain in information cost more than enough to compensate for their loss in value. This may hold even if M is differentially private. As a specific instantiation, we prove that if a mechanism M approximately solves the 1-facility location problem over any metric space and simultaneously releases a synopsis, then M cannot be truthful. See Section 5.2 for formal statements.

2.4 Comparison with related work

Since the original version of this manuscript appeared in January 2011 [35], in the following discussion we separate "previous work" and "subsequent work" according to that date.

2.4.1 Previous work.

There has already been a fruitful interaction between game theory and the study of differential privacy. The first work combining the two is the "Exponential Mechanism" of McSherry and Talwar [27], which achieves privacy, *approximate* truthfulness, and efficiency. Approximate truthfulness says that the amount any player can gain by announcing a false type is small (but possibly positive). Approximate truthfulness is in fact a consequence of differential privacy, which says that any one player's type can only affect the output distribution of the mechanism by a little. McSherry and Talwar [27] interpret approximate truthfulness to mean that the player might as well announce his type honestly, since he has little incentive to deviate. They also show that their mechanism achieves good efficiency (assuming the players behave truthfully).

Unfortunately, the approach suffers from the following two drawbacks, first observed by Nissim et al. [31]. First, in the Exponential Mechanism, *all* possible strategies of any individual player lead to approximately the same output distribution of the mechanism. If one accepts the interpretation that players are indifferent to small changes in their utility, then intuitively privacy may even motivate players to *lie*: players are indifferent to the small amount of utility from the game outcome that is lost by lying, but they would prefer to lie because that would reduce the amount of information leaked by the mechanism about their private type. (We will formalize this concern in Section 5, see also the discussion in Section 2.3.) Thus, it seems that approximate truthfulness cannot substitute for standard truthfulness when privacy is desired.

Second, Nissim et al. [31] showed that the Exponential Mechanism is *not* truthful for a game based on digital goods auctions, and therefore the relaxation to approximate truthfulness is inherently necessary. (As we mentioned, we show that Exponential Mechanism is not truthful also for the $\mathsf{LINE\text{-}1\text{-}FAC}$ game.)

Nissim et al. [31] go on to show that by combining the Exponential Mechanism with a "Gap Mechanism" that incentivizes honesty, one can get a fully truthful mechanism. They apply this mechanism to give a truthful and approximately efficient mechanism for the k-facility problem on a line, which is a more general version of the 1-facility problem on a line that we will study in this paper. Unfortunately, their mechanism is not differentially private, because the Gap Mechanism relies on constraining the post-actions of the players, and this constraint reveals the types of the players in a very non-private way.

Ghosh and Roth [18] consider a question that is related but orthogonal to our work: they consider *verified* databases where each player has private information that a database owner wants to gather. Their mechanism allows each player i to declare a pricing function $c_i : \mathbb{R} \to \mathbb{R}$ such that $c_i(\varepsilon)$ represents the minimal payment that player i would require to submit his information to a database that is then published via an ε-differentially private sanitization mechanism. Their mechanism uses these pricing functions to compute a value of ε and payment values that are paid out to each player such that enough of them are incentivized to participate to make the outcome of the sanitizer accurate. More follow-up works in this direction have appeared recently [32, 17, 25]. The main difference with our model is that in this line of work, it is assumed that the players' private information will be accurately reported, and they may *only* lie about how much they value their privacy. In contrast, our model focuses on deviations resulting from players reporting false private information, but does not explicitly consider players' lying about their valuation of their privacy.

Feigenbaum et al. [14] study how to keep information private even from the database owner, *i.e.* before running sanitization. We do not study this problem here and treat the database owner as a trusted party. We note however that, using standard cryptographic assumptions and protocols, one can replace a trusted database owner by a secure multiparty computation among the individuals.

2.4.2 Subsequent work

Subsequent to the initial version of this manuscript, Chen et al. [4] showed that it is possible to build truthful mechanisms even when the players' utilities explicitly take into account the cost of information leaked, thus answering affirmatively one of the open questions posed in the initial version of this manuscript. Their result holds for a general class of information costs. Some of their mechanisms are inspired by the TE-to-PTE transformation (Theorem 4.3) presented in this paper.

Huang and Kannan [23] resolve another problem posed in the original version of this manuscript: they show that one can alter the VCG mechanism to give a PTE mechanism for *any* social welfare game. This bypasses the restriction of our transformation to small type spaces. We note that their result is nevertheless incomparable to ours in certain respects, because it inherently uses payments and so cannot be used to construct moneyless mechanisms, and their mechanism is not necessarily computationally efficient (although for certain specific games they acheive computational efficiency).

Nissim et al. [30] show that under certain assumptions about the distribution of information costs of a population and assuming that the mechanism can restrict the reactions of individuals, one can build truthful mechanisms that take into account the players' information costs.

3 Preliminaries

We let $[q] = \{1, \ldots, q\}$. We identify sets S with the uniform distribution over the set. For a distribution X, we write $x \leftarrow X$ to denote a random sample from that distribution. We let underlined variables $\underline{x}$ denote vectors and x_i the i'th coordinate in the vector. For $\underline{x} \in \mathbb{R}^n$, we let $\|\underline{x}\|_1 = \sum_{i=1}^n |x_i|$ denote the ℓ_1 norm and $\|\underline{x}\|_\infty = \max_i |x_i|$ the ℓ_∞ norm.

Games A game with n players consists of the type space Q, a space S of possible outcomes of the game, a valuation function $v : Q \times S \to [-1, 1]$, and a global utility function $w : Q^n \times S \to \mathbb{R}$. The valuation function v determines the private utilities of the players. When Q is finite, we let $|Q| = q$. For $i \in [n]$, for $t_1, \ldots, t_n$, we let $\underline{t}^{-i}$ denote the vector with $n-1$ entries given by $t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n$. Define $\underline{h}(\underline{t}) \in \mathbb{Z}^q$ the histogram of the input, where $h_j = |\{i \mid t_i = j\}|$. We assume *w.l.o.g.* that the type space has a special $\perp$ type, and any player that declares this special type is ignored. (Namely, an input with n players, k of whom have type $\perp$, is treated as an input to the game with $n - k$ players with those k players removed.)

We define the following properties of games:

1. A game is *anonymous* if it has an anonymous global utility function: let w be its global utility function, then there is a function w' such that for all $\underline{t}, s$, it holds that $w(\underline{t}, s) = w'(\underline{h}(\underline{t}), s)$. In such cases, we

abuse notation and write $w(\underline{h}, s)$ to mean $w(\underline{t}, s)$ where $\underline{h} = \underline{h}(\underline{t})$.

2. An anonymous global utility function is *insensitive* if for all integers k, $\|\underline{h} - \underline{h}'\|_1 \leq k$ implies that for all s, $|w(\underline{h}, s) - w(\underline{h}', s)| \leq k$.
3. The *social welfare* is $w(\underline{t}, s) = \sum_{i=1}^n v(t_i, s)$. It is anonymous and insensitive.

Mechanisms A mechanism M is a (randomized) function that takes types $t_1, \ldots, t_n$ for all of the players and samples an outcome $(s, p) \leftarrow M(t_1, \ldots, t_n)$ where $s \in S$ is the game outcome and $p : Q \to \mathbb{R}$ is a function determining payments. Namely, any player with type t must pay $p(t)$ to the mechanism (since we only consider anonymous games, payments only depend on the type and not the identity of the player). Each player tries to maximize $v(t_i, s) - p(t_i)$ where (s, p) is the outcome of the game.

We say that M is *moneyless* if $p \equiv 0$ for all inputs and random coins. We are concerned with asymptotic analysis, so M must be be able to handle any number of players n.

A player strategy σ is a function mapping the type space Q to distributions over Q. If a player uses strategy σ, this means given a true type t, the player samples from $t' \leftarrow \sigma(t)$ and reports t' to the mechanism.

1. We say a mechanism M is *truthful* if for every player $i \in [n]$, and for all $\underline{t}$ and all strategies σ, it holds that $\mathbb{E}_{(s,p)\leftarrow M(\underline{t})} M[v(t_i, s) - p(t_i)] \geq \mathbb{E}_{t'_i \leftarrow \sigma(t_i), (s,p) \leftarrow M(\underline{t}^{-i}, t'_i)}[v(t_i, s) - p(t'_i)]$.[3]
2. We say a mechanism is $\delta(n)$-*efficient* if for all inputs $\underline{t}$ on n players, it holds that $\mathbb{E}_M[w(\underline{t}, M(\underline{t}))] \geq \max_{s \in S} w(\underline{t}, s) - \delta(n)$.

Observe that our definition of efficiency allows for an *additive* error. In contrast, most work in the mechanism design literature on approximate mechanisms deal with *multiplicative* error. However, additive error is more suitable when working with differential privacy.

Differential privacy A mechanism M is (ε, η)-*differentially private* if for all $i \in [n]$, all $\underline{t} \in Q^n$, all $t' \in Q$, and all subsets U of the output space of M, it holds that $\Pr[M(\underline{t}) \in U] \leq e^{\varepsilon} \Pr[M(\underline{t}^{-i}, t') \in U] + \eta$. Typically we think of $\varepsilon > 0$ being a small constant and η as being $o(1)$, preferably negligible.

Definition 3.1 (PTE mechanisms)**.** A mechanism M for a game define by private valuation functions $v : Q \to [-1, 1]$ and global utility function $w : Q^n \to \mathbb{R}$ is $(\varepsilon, \eta, \delta)$-PTE if it is (ε, η)-differentially private, truthful (with respect to private valuations v), and δ-efficient (with respect to global utility w).

Useful distributions We will frequently write $\alpha = e^{-\varepsilon}$. Let $\mathcal{G}_\varepsilon$ denote the geometric distribution over the integers, with probability mass function $f(x) = \frac{1-\alpha}{1+\alpha}\alpha^{|x|}$. Let $\mathcal{H}_{\varepsilon,\tau,q}$ be the following distribution: sample $\underline{\zeta} \leftarrow \mathcal{G}_\varepsilon^q$. If $\|\underline{\zeta}\|_\infty > \tau$, output 0, otherwise output $\underline{\zeta}$.

It is straightforward to calculate that for $\zeta \leftarrow \mathcal{G}_\varepsilon$:

$$\Pr[|\zeta| \geq \tau] = \frac{2\alpha^\tau}{1 + \alpha} \tag{3.1}$$

Lemma 3.2. *For all $i, j \in [q]$ and $U \subseteq \mathbb{Z}^q$, for $\underline{\zeta}'$ sampled from $\mathcal{H}_{\varepsilon,\tau,q}$ it holds that:*

$$\Pr[\underline{\zeta}' \in U] \leq e^{2\varepsilon} \Pr[e_i - e_j + \underline{\zeta}' \in U] + \tfrac{2q\alpha^\tau}{1+\alpha}$$

where e_i denotes the i'th standard basis vector.

For completeness we provide a proof in Section B.1.

Negative binomial distribution The negative binomial distribution $\mathcal{NB}_{q,\varepsilon}$ is defined using the probability mass function

$$f(x) = \binom{x + q - 1}{q - 1} \cdot (1 - e^{-\varepsilon})^q e^{-\varepsilon x}$$

[3] Observe that the value is calculated according to the true type t_i on both sides, but the payment is calculated according to the declared type (t_i in one case, t'_i in the other). This is because the payment is what the mechanism asks the player to play, and this is a function of the *declared* type.

Fact 3.3 (*e.g.* [34])**.** *We note the following:*

1. *The sum of q two-sided random geometric variables $\sum_{j=1}^{q} \zeta_j$, where each ζ_j is distributed according to $\mathcal{G}_\varepsilon$, is distributed identically to $Y - Y'$ where both Y, Y' are independent and identically distributed according to the $\mathcal{NB}_{q,\varepsilon}$.*[4]

2. *Suppose Y is distributed according to $\mathcal{NB}_{q,\varepsilon}$. Then $\Pr[Y \geq t] = \Pr[Z \leq q]$ where Z is a binomial random variable for an experiment with $q + t$ trials and success probability $(1 - e^{-\varepsilon})$ for each trial.*

4 PTE Mechanisms

4.1 The Exponential Mechanism not truthful for LINE-1-FAC

Definition 4.1. The LINE-1-FAC game is defined as follows. The player types are $t_i \in [0, 1]$. The outcome of the game is a point $s \in [0, 1]$. The utility function is $v(t, s) = -|t - s|$ and the global utility is the social welfare: $w(\underline{t}, s) = \sum_{i\in[n]} v(t_i, s) = -\sum_{i\in[n]} |t_i - s|$.

The moneyless mechanism that outputs the median (breaking ties say by picking the left median point) of the $\{t_1, \ldots, t_n\}$ is truthful and achieves optimal social welfare for this game [28].

We now prove Theorem 2.1, the fact that the Exponential Mechanism is not truthful for LINE-1-FAC.

of Theorem 2.1. For $n = 2$, set $t_1 = 0, t_2 = 2/3$. (This can easily be extended to an arbitrary even number of players n by placing $n/2$ players at 0 and $(n/2 - 1)$ players at 1, and one player at 2/3.)

Claim 4.2. *For all $\varepsilon > 0$, if player 2 declares 1 then his utility under the Exponential Mechanism is $-5/18$, while if he declares $2/3$ then his utility is strictly less than $-5/18$.*

of Claim 4.2. The Exponential Mechanism M_{Exp} generates an output s according to the density function $f(s) = \frac{e^{\varepsilon w(\underline{t}, s)}}{\int_0^1 e^{\varepsilon w(\underline{t}, s)} ds}$. One can compute that the expected utility of player 2 if he reports 1 is $-\int_0^1 |2/3 - s| ds = -5/18$, since in this case there are exactly half the players at 0 and at 1 and so the social welfare function (and hence the mechanism's output) is uniform over $[0, 1]$.

If player 2 is truthful, then the social welfare equals $w(\underline{t}, s) = -s - |2/3 - s|$. The idea is that this welfare decreases as s increases past 2/3, and so the Exponential Mechanism will give lower weight to the points in $[2/3, 1]$. This will hurt player 2 and he will get worse utility than if he declared that he was at 1.

We formalize this and analyze the Exponential Mechanism for a truthful player 2. The expected utility of player 2 is:

$$V \overset{def}{=} \mathbb{E}_{s \leftarrow M_{\mathsf{Exp}}(0,2/3)}[v(2/3, s)] = \frac{-\int_0^1 e^{-\varepsilon(s+|2/3-s|)}|2/3 - s| ds}{\int_0^1 e^{-\varepsilon(s+|2/3-s|)} ds}$$

We claim that for all $\varepsilon > 0$, it holds that $V < -5/18$. This is equivalent to proving

$$0 < -V - 5/18 = \frac{\int_0^1 e^{-\varepsilon(s+|2/3-s|)} \cdot (|2/3 - s| - 5/18) ds}{\int_0^1 e^{-\varepsilon(s+|2/3-s|)} ds}$$

Observe that the denominator is positive, so it suffices to prove that the numerator is positive for all $\varepsilon > 0$. Evaluating the integral we obtain that

$$\int_0^1 e^{-\varepsilon(s+|2/3-s|)} \cdot (|2/3 - s| - 5/18) ds = e^{-2\varepsilon/3} \left(\frac{1}{27} + \frac{9 - 5\varepsilon - e^{-2\varepsilon/3}(\varepsilon + 9)}{36\varepsilon^2} \right) \quad (4.1)$$

Since $e^{-2\varepsilon/3} > 0$ so we can multiply the LHS of Equation 4.1 by $36\varepsilon^2 e^{2\varepsilon/3}$ and simplify, and it suffices to prove that

$$\tfrac{4\varepsilon^2}{3} + 9 - 5\varepsilon - e^{-2\varepsilon/3}(\varepsilon + 9) > 0 \quad (4.2)$$

[4] This follows from the fact that a two-sided geometric random variable is identical to the difference between two one-sided geometric random variables, and the sum of one-sided geometric random variables is a negative binomial.

```
Input: types $t_1, \dots, t_n \in [q]$. Auxiliary input: $\varepsilon, \eta$ privacy parameters. Set $\tau = O(\log(q/\eta)/\varepsilon)$.
  1. Sample $\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}$
  2. Construct $\underline{h}' = \underline{h} + \underline{\zeta} + \tau \cdot \underline{1}$, where $\underline{1}$ is the all 1 vector.
  3. Output $M(\underline{h}')$.
```

Algorithm 4.4. PTE mechanism based on a truthful and efficient mechanism M.

For large $\varepsilon \gg 0$ the RHS of Equation 4.2 is dominated by the quadratic term. For instance it is easy to check that for all $\varepsilon > 1$, it holds that $e^{-2\varepsilon/3} < 0.52$ and so:

$$\begin{aligned}
\tfrac{4\varepsilon^2}{3} + 9 - 5\varepsilon - e^{-2\varepsilon/3}(\varepsilon + 9) &> \tfrac{4\varepsilon^2}{3} + 9 - 5\varepsilon - 0.52(\varepsilon + 9) \qquad (4.3)\\
&> \tfrac{4\varepsilon^2}{3} - 4.48\varepsilon + 4.32 \qquad (4.4)\\
&> 0 \qquad (4.5)
\end{aligned}$$

where the final inequality can be deduced by the fact that the quadratic polynomial in Equation 4.4 has no real roots and is non-negative. Therefore we can restrict our attention to the case $\varepsilon \leq 1$. Using the Taylor expansion of the exponential, we know that $e^{-2\varepsilon/3} \leq 1 - \frac{2\varepsilon}{3} + \frac{2\varepsilon^2}{9} - \frac{4\varepsilon^3}{81} + \frac{2\varepsilon^4}{243}$ for $\varepsilon \leq 1$, therefore Equation 4.2 can be rewritten as:

$$\begin{aligned}
&\tfrac{4\varepsilon^2}{3} + 9 - 5\varepsilon - e^{-2\varepsilon/3}(\varepsilon + 9)\\
&\quad \geq \tfrac{4\varepsilon^2}{3} + 9 - 5\varepsilon - (1 - \tfrac{2\varepsilon}{3} + \tfrac{2\varepsilon^2}{9} - \tfrac{4\varepsilon^3}{81} + \tfrac{2\varepsilon^4}{243})(\varepsilon + 9)\\
&\quad = \tfrac{4\varepsilon^2}{3} + 9 - 5\varepsilon - \varepsilon - 9 + \tfrac{2\varepsilon^2}{3} + 6\varepsilon - \tfrac{2\varepsilon^3}{9} - 2\varepsilon^2\\
&\qquad + \tfrac{4\varepsilon^4}{81} + \tfrac{4\varepsilon^3}{9} - \tfrac{2\varepsilon^5}{243} - \tfrac{2\varepsilon^4}{27}\\
&\quad = \tfrac{2\varepsilon^3}{9} - \tfrac{2\varepsilon^4}{81} - \tfrac{2\varepsilon^5}{243}\\
&\quad > \varepsilon^3(\tfrac{2}{9} - \tfrac{2}{27} - \tfrac{2}{27})\\
&\quad > 0
\end{aligned}$$

where in the last two lines we use the fact that $\varepsilon \leq 1$. ■

■

4.2 Transformation from TE to PTE

We now exhibit a generic transformation that converts a truthful and efficient mechanism M into a PTE mechanism M'. We assume M satisfies the following: on input $\underline{t}$, M computes its output depending only on $\underline{h} = \underline{h}(t)$ (and never looks at the entries of $\underline{t}$ individually). (This is without loss of generality for anonymous games, see Appendix A.)

Theorem 4.3. *Let G be a game with type space of size q and whose global utility function is anonymous and insensitive. Suppose G has a truthful and δ-efficient histogram mechanism M and that δ is non-decreasing.*[5]

Then for all $\varepsilon, \eta > 0$, G also has a $(2\varepsilon, \eta, \delta')$-PTE mechanism M', where $\delta'(n) = \delta(n + O(q \log(q/\eta)/\varepsilon)) + O(q \log(q/\eta)/\varepsilon)$. M' is given by Algorithm 4.4. Note that if M is computationally efficient, so is M', and if M is moneyless, so is M'.

of Theorem 4.3. Let M be the truthful and δ-efficient mechanism. By Equation 3.1 and the union bound, we can set $\tau = O(\log(q/\eta)/\varepsilon)$ such that

$$\Pr_{\underline{\zeta} \leftarrow \mathcal{G}_\varepsilon^q}[\|\zeta\|_\infty > \tau] < \frac{2qe^{-\varepsilon\tau}}{1 + e^{-\varepsilon}} = \eta \qquad (4.6)$$

[5] This is a technical condition to simplify the expression for δ'; for any interesting δ we can artificially increase it as some points to make it satisfy this property.

In Algorithm 4.4 we construct a mechanism M' (using τ set as just mentioned) that is private, truthful, and δ'-approximately efficient. We prove that M' is PTE:

Truthfulness Fix an input $\underline{t}$. We claim that for every choice of $\underline{\zeta}$, the mechanism is truthful. We run M on the histogram $\underline{h}' = \underline{h} + \underline{\zeta} + \tau \cdot \underline{1}$. Because $\underline{\zeta}$ is sampled from $\mathcal{H}_{\varepsilon,\tau,q}$, it holds for every $t \in Q$ that $h'_t \geq h_t \geq 0$. Suppose that, for this fixing of $\underline{\zeta}$, there is a deviation that benefits some player, *i.e.* there exists i and σ such that

$$\mathbb{E}_{(s,p)\leftarrow M(\underline{h}')}[v(t_i, s) - p(t_i)] < \mathbb{E}_{t' \leftarrow \sigma(t_i),(s,p)\leftarrow M(\underline{h}' - e_{t_i} + e_{t'})}[v(t_i, s) - p(t')]$$

where $e_t \in \mathbb{Z}^q$ is the t'th standard basis vector. Then this is also a deviation on the input $\underline{h}'$ for the original M, which contradicts the fact that M is truthful.

Efficiency Let $\underline{\zeta}$ be sampled from $\mathcal{H}_{\varepsilon,\tau,q}$ and observe that $\|\underline{\zeta}\|_\infty \leq \tau$. For any input $\underline{t}$ define $\underline{h} = \underline{h}(\underline{t})$ and the modified histogram $\underline{h}' = \underline{h}(\underline{t}) + \underline{\zeta} + \tau \cdot \underline{1}$. Observe that $|\underline{h} - \underline{h}'|_1 \leq 2q\tau$. By the insensitivity of w, it holds that for all possible outcomes of the game s', it holds that $\left|w(\underline{h}', s') - w(\underline{h}, s')\right| \leq 2q\tau$. Therefore it holds that:

$$\begin{aligned}
\mathbb{E}_M[w(\underline{h}, M(\underline{h}'))] &\geq \mathbb{E}_M[w(\underline{h}', M(\underline{h}'))] - 2q\tau \qquad (4.7)\\
&\geq \max_{s'} w(\underline{h}', s') - 2q\tau - \delta(n + 2q\tau) \qquad (4.8)\\
&\geq w(\underline{h}', s_0) - 2q\tau - \delta(n + 2q\tau) \qquad (4.9)\\
&\geq \max_{s} w(\underline{h}, s) - 4q\tau - \delta(n + 2q\tau) \qquad (4.10)
\end{aligned}$$

In Equation 4.8 we use the efficiency of M, the fact that $\underline{h}'$ corresponds to a game with at most $n + 2q\tau$ players, and that $\delta(n) \in [0, 2n]$ is non-decreasing. In Equation 4.10 s_0 denotes the outcome that maximizes $w(\underline{h}, s)$.

Applying Equation 4.10 to the expected utility of M', one obtains the following:

$$\begin{aligned}
\mathbb{E}_{M'}[w(\underline{t}, M'(\underline{h}))] &= \mathbb{E}_{\underline{\zeta}\leftarrow\mathcal{H}_{\varepsilon,\tau,q},M}[w(\underline{t}, M(\underline{h}'))]\\
&\geq \max_{s} w(\underline{t}, s) - \delta(n + 2q\tau) - 4q\tau
\end{aligned}$$

Privacy Observe that Algorithm 4.4 is just perturbing the histogram according to the distribution $\mathcal{H}_{\varepsilon,\tau,q}$. Therefore, from Lemma 3.2, we know that for any adjacent $\underline{h}, \underline{h}^*$, it holds that for all subsets $U \subseteq \mathbb{Z}^q$ that

$$\Pr[\underline{h} + \underline{\zeta}' \in U] \leq e^{2\varepsilon} \Pr[\underline{h}^* + \underline{\zeta}' \in U] + \tfrac{2qe^{-\varepsilon\tau}}{1+e^{-\varepsilon}}$$

Since by our choice of τ we know that $\frac{2qe^{-\varepsilon\tau}}{1+e^{-\varepsilon}} \leq \eta$, this shows the mechanism is $(2\varepsilon, \eta)$-differentially private. ■

Note that one popular alternative way of perturbing the histogram, by adding noise according to $\mathcal{G}_\varepsilon$ to each bin of the histogram and then truncating negative bins to 0, does not seem to guarantee truthfulness. The problem is that the truncation process is non-linear, and so it is unclear how to prove that a deviation in the perturbed game implies a deviation in the original game.

4.3 LINE-1-FAC has a PTE mechanism

We show that although the Exponential Mechanism is not truthful for LINE-1-FAC (Definition 4.1), there does exist a mechanism that is truthful, differentially private, and approximately efficient. This mechanism is given in Algorithm 4.10. The idea is to reduce LINE-1-FAC to a game D-L1F ("discrete one-facility on a line") with a small type space, and then to give a private, truthful, and efficient mechanism for D-L1F using Theorem 4.3.

```
Input: histogram of player types $\underline{h}$. Let $n = \sum_{j=1}^{q} h_j$.
  1. Output the minimal $s \geq 1$ such that $\sum_{j=1}^{s} h_j \geq \frac{n}{2}$.
```

Algorithm 4.6. Truthful and efficient mechanism for D-L1F

4.3.1 The D-L1F game

Definition 4.5. The D-L1F$_\gamma$ game is defined as follows. Assume that $\gamma > 0$ is such that $q = 1/\gamma + 1$ is an integer. The player types are $t_i \in [q]$. Output of mechanism is $s \in [q]$. Utility function is $v(t, s) = -\gamma|t - s|$. Global utility is social welfare: $w(\underline{t}, s) = \sum_{i=1}^{n} v(t_i, s)$.

Theorem 4.7. *There is a truthful and perfectly efficient mechanism for D-L1F$_\gamma$.*

The mechanism is given in Algorithm 4.6. We may apply Theorem 4.3, we obtain:

Corollary 4.8. *For all $\varepsilon, \eta > 0$, D-L1F$_\gamma$ has a $(\varepsilon, \eta, \delta)$-PTE mechanism for $\delta = O(\log(\frac{1}{\gamma\eta})/(\gamma\varepsilon))$.*

The proof of Theorem 4.7 is a special case of the proof that the median mechanism is truthful and efficient for single-peaked preferences. For completeness, we give a proof in Section B.2.

4.3.2 Using D-L1F$_\gamma$ to give a PTE mechanism for LINE-1-FAC

Theorem 4.9. *For any $\gamma, \varepsilon, \eta > 0$, the mechanism of Algorithm 4.10 is $(2\varepsilon, \eta)$-differentially private, truthful, and δ-efficient for the LINE-1-FAC game, where $\delta = n\gamma + O(\frac{1}{\gamma\varepsilon}\log(\frac{1}{\gamma\eta}))$.*

As an example setting, pick $\gamma = n^{-1/2}$, which implies $\delta = O(n^{1/2}\log(n/\eta)/\varepsilon)$.

of Theorem 4.9. $(2\varepsilon, \eta)$-differential privacy follows immediately from Theorem 4.3. Efficiency is also straightforward, because the overall error is at most the discretization error, which is bounded by γ for each player, plus the error from M', which is bounded by $O(\frac{1}{\gamma\varepsilon}\log(\frac{1}{\gamma\eta}))$.

Truthfulness. Truthfulness must be argued more carefully because the rounding process might cause unexpected problems. By symmetry, it suffices to consider player 1.

Fix $t_1, \underline{t}^{-1}$, and t^*. We will show that the player can gain no utility from declaring t^* when his actual type is t_1. Fix any choice of random coins $\underline{\zeta}$ used by M'. Let $\underline{h}' = \underline{h}(\underline{t}) + \underline{\zeta} + \tau \cdot \underline{1}$ and $\underline{h}^* = \underline{h}(\underline{t}^{-1}, t^*) + \underline{\zeta} + \tau \cdot \underline{1}$. Let $s = M(\underline{h}')$ and $s^* = M(\underline{h}^*)$. From the proof of truthfulness of Algorithm 4.6, we observe that M has the following property (which is stronger than truthfulness): either $s = s^*$ or $|\hat{t}_1 - s| \leq |\hat{t}_1 - s^*| - 1$. (Namely, it is impossible for $s^* \neq s$ and yet $|\hat{t}_1 - s| = |\hat{t}_1 - s^*|$.)

In the case where $s = s^*$ then there is clearly no advantage to lying, so suppose $|\hat{t}_1 - s| \leq |\hat{t}_1 - s^*| - 1$. Observe that the rounding process guarantees that for all $t \in [0, 1]$, $|t - \gamma\hat{t}| \leq \gamma/2$. Therefore we may write

$$\begin{aligned} |t_1 - \gamma s| &\leq \gamma|\hat{t}_1 - s| + \gamma/2 \\ &\leq \gamma(|\hat{t}_1 - s^*| - 1) + \gamma/2 \\ &\leq |t_1 - \gamma s^*| \end{aligned}$$

This again implies that there is no advantage to declaring t^*, and so since for every choice of $\underline{\zeta}$ the mechanism is truthful, it follows that the overall mechanism is truthful. ■

Input: player types $t_1, \dots, t_n \in [0, 1]$. Assume for simplicity that $1/\gamma$ is an integer.

1. Discretize $[0, 1]$ into $q = 1/\gamma + 1$ intervals: $[0, \gamma/2), [\gamma/2, (1 + 1/2)\gamma), \dots, [(j - 1/2)\gamma, (j + 1/2)\gamma), \dots, [(q - 1/2)\gamma, 1]$.
2. Assign player i the proxy type $\hat{t}_i = j$ such that t_i falls into the j'th interval.
3. Let M denote the mechanism for $\mathsf{D\text{-}L1F}_\gamma$ given by Theorem 4.7. Let M' be the corresponding $(2\varepsilon, \eta)$-differentially private, truthful, and efficient mechanism, given by Corollary 4.8. Run $M'(\hat{t}_1, \dots, \hat{t}_n)$ to obtain $s \in [q]$.
4. Output γs.

Algorithm 4.10. $(2\varepsilon, \eta)$-differentially private, truthful, and efficient mechanism for $\mathsf{LINE\text{-}1\text{-}FAC}$.

4.4 Application to the VCG mechanism

Corollary 2.4, stating that all social welfare games with small type space have a PTE mechanism, follows from applying Theorem 4.3 to the VCG mechanism (which is truthful and perfectly efficient for all social welfare games).

Observe that the privacy holds with respect to the outcome of the game s and the payment function p that maps types to payments. In particular, we assume that the payments of individual players are not revealed publicly. Clearly, privacy would not hold if outside observers learned that, say, player i made payment $p(t_i)$. This is unavoidable, but it seems reasonable that the actual payment of each player can be kept secret *e.g.* by transmitting the payment using cryptographic techniques or some other secure channel. One can also ensure that the player and curator do not try to underpay/overcharge by using proofs that the mechanism was computed correctly and that amount paid is correct by using traditional cryptographic techniques (*e.g.* NIZK proofs).

5 The value of privacy

We now explore a model that assigns non-negative costs to the information leaked by mechanisms about the private types of the players. As argued in the introduction, we assume the information cost to player i depends on the mechanism M, the strategy σ used by player i, and the types of all players $\underline{t}$. In addition we will allow the information cost with respect to a small approximation factor $\eta \geq 0$, which we motivate below. Therefore, the information cost is a function of the form $\mathsf{IC}^\eta_M(\sigma, \underline{t}, i)$. We will sometimes write IC_M to denote IC^0_M. Our formal assumptions about the information cost are:

Assumption 5.1. We make the following assumptions about the η-approximate information cost:

Players can hide their data: if σ is independent of its input (*i.e.* $\sigma(t) = \sigma(t')$ for all t, t'), then for all $i, \underline{t}$ it holds that $\mathsf{IC}^\eta_M(\sigma, \underline{t}, i) = 0$.

Cost reflects differential privacy: for the truthful (*i.e.* identity) strategy Id:

$$\mathsf{IC}^\eta_M(\mathsf{Id}, \underline{t}, i) \geq \Omega\left(\max_{t' \in Q} \max_B \log \frac{\Pr[M(\underline{t}^{-i}, t) \in B] - \eta}{\Pr[M(\underline{t}^{-i}, t') \in B](1 - \eta)}\right) \tag{5.1}$$

where the maximum of B is taken over all subsets such that $\Pr[M(\underline{t}^{-i}, t) \in B] > \eta$.

The intuition behind these assumptions was discussed in Section 2.3.1. The formal definition of the assumption that cost reflects differential privacy (Equation 5.1) intuitively says that the cost is at least how much more likely the true type t is than any other type t' after having observed the output of the mechanism. This is precisely the kind of difference in probability captured by differential privacy. In fact, one natural definition of information cost in the case where we have no further application-specific criteria is precisely the RHS of Equation 5.1 (see Appendix C).

We now mention the role of η. When we study (ε,η)-differentially private mechanisms, we will set η in the assumption to roughly match the η in (ε,η)-differential privacy. This is intuitively natural, since without this η the mechanism's output may have small fluctuations in probability that create large (but intuitively unmeaningful) information costs. For ε-differential privacy, we set $\eta = 0$.

To combine the information cost with the value derived by players from a game, we use the following

$$u_i(\sigma, \underline{t}) = \mathbb{E}_{t' \leftarrow \sigma(t), s \leftarrow M(\underline{t}^{-i}, t')}[v(t_i, s)] - \nu_i \mathsf{IC}_M(\sigma, \underline{t}, i) \tag{5.2}$$

Here, we have weighted the information cost with a factor ν_i that expresses how much the individual values his privacy relative to the value of the game.

Remark 5.2. In Equation 5.2, ν_i models the weight of player i's privacy, and for simplicity we assume that the ν_i are fixed and known (Ghosh and Roth [18] study mechanisms where these valuations are private). Intuitively $1/\nu_i$ represents how many bits must be leaked for player i to lose a constant amount of utility. $\nu_i = \Omega(1/\log|Q|)$ is a realistic setting, and would mean a constant cost is incurred if a constant *fraction* of bits is leaked. Signficantly smaller values of ν_i would model a situation in which the player assigns significant cost *only* when his type is essentially *completely* revealed by the mechanism. In particular, we may safely assume that the weights to satisfy $\nu_i = 2^{-o(n)}$, since otherwise the amount of utility that the player places on privacy is so small that explicitly modelling the value of privacy loses relevance.

5.1 Releasing histograms for LINE-1-FAC

Recall that Corollary 2.3 says that by applying our transformation (Theorem 4.3) to a discretized version of the median mechanism for the LINE-1-FAC game, we can give a PTE mechanism for the LINE-1-FAC game. An explicit description of this mechanism is given in Algorithm 4.10, let us call this mechanism M'.

Here we show that, on certain databases, no single player can heavily influence the outcome of M'. $(2\varepsilon, \eta)$-differential privacy implies that being untruthful can hurt the value of a player by at most $2\varepsilon + \eta$. It turns out that for this particular mechanism M', there exist inputs for which the loss is much smaller.

Theorem 5.3. *Fix any $\varepsilon > 0$ and any $\eta > 0$ such that $\eta = 2^{-o(\sqrt{n}/\log n)}$ (*i.e. *η is not too small). Then for n large enough, for all i, there exists $\underline{t}^{-i} \in [0,1]^{n-1}$ such that for all $t_i, t'_i \in [0,1]$, it holds that:*

$$\mathbb{E}_{s \leftarrow M'(\underline{t})}[v(t_i, s)] - \mathbb{E}_{s \leftarrow M'(\underline{t}^{-i}, t'_i)}[v(t_i, s)] \leq e^{-(1-e^{-\varepsilon})^2 n} \tag{5.3}$$

Proof. To simplify our notation, suppose that the number of players is $n+1$ rather than n. By our choice of γ, Algorithm 4.10 divides $[0,1]$ into $q = n^{1/2}$ intervals.

We study the case $i = 1$, as symmetry will imply that the same argument works for all choices of i. The setting of player types $\underline{t}^{-1}$ is simply to put n players at location 0.

Recall that M' functions by building a histogram of the players' locations, generating noise $\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}$ where $\tau = O(\log(q/\eta)/\varepsilon)$, adding $\underline{\zeta} + \tau \cdot \underline{1}$ to the histogram, and then running the deterministic mechanism M of Algorithm 4.6. Notice that for our choice of $\underline{t}^{-1}$, there is no rounding necessary. Let us rewrite the LHS of Equation 5.3 as:

$$\mathbb{E}_{\underline{\zeta}}[v(t_1, M(\underline{h}(\underline{t}) + \underline{\zeta} + \tau \cdot \underline{1})) - v(t_1, M(\underline{h}(\underline{t}^{-1}, t'_1) + \underline{\zeta} + \tau \cdot \underline{1}))] \tag{5.4}$$

The main observation is that for most choices of $\underline{\zeta}$, M gives the same output regardless of what player 1 declares as its value. To state this more formally, define

$$\underline{h}^{-1} = \underline{h}(\underline{t}^{-1}) + \underline{\zeta} + \tau \cdot \underline{1} \tag{5.5}$$

$$n' = (n+1) + \sum_{j=1}^{q} \zeta_j + q\tau \tag{5.6}$$

The value in Equation 5.4 is upper-bounded by the probability over ζ that $M(\underline{h}(\underline{t}) + \underline{\zeta} + \tau \cdot \underline{1}) \neq M(\underline{h}(\underline{t}^{-1}, t'_1) + \underline{\zeta} + \tau \cdot \underline{1})$. Call this event B. B occurs only when there exists $k \in [q]$ such that $\sum_{j=1}^{k} (\underline{h}^{-1})_j = \lceil n'/2 - 1 \rceil$.

Claim 5.4. $\Pr[B] \le e^{-(1-\alpha)^2 n}$

Since we argued above that $\mathbb{E}_{\underline{\zeta}}[v(t_1, M(\underline{h}(\underline{t}) + \underline{\zeta} + \tau \cdot \underline{1})) - v(t_1, M(\underline{h}(\underline{t}^{-1}, t_1') + \underline{\zeta} + \tau \cdot \underline{1}))] \le \Pr_{\underline{\zeta}}[B]$, this claim implies the theorem. ■

of Claim 5.4. Observe that $\Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,q,\tau}}[B]$ can be written as:

$$\Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,q,\tau}}\left[\|\underline{\zeta}\|_\infty \le \tau \wedge \exists k,\ \sum_{j=1}^{k} (\underline{h}^{-1})_j = \lceil n'/2 - 1 \rceil\right] = \Pr_{\underline{\zeta} \leftarrow \mathcal{G}_\varepsilon^q}\left[\|\underline{\zeta}\|_\infty \le \tau \wedge \exists k, \sum_{j=1}^{k} (\underline{h}^{-1})_j = \lceil n'/2 - 1 \rceil\right] \tag{5.7}$$

$$\le \Pr_{\underline{\zeta} \leftarrow \mathcal{G}_\varepsilon^q}\left[\exists k,\ \sum_{j=1}^{k} (\underline{h}^{-1})_j = \lceil n'/2 - 1 \rceil\right] \tag{5.8}$$

$$\le \sum_{k=1}^{q} \Pr_{\underline{\zeta} \leftarrow \mathcal{G}_\varepsilon^q}\left[\sum_{j=1}^{k} (\underline{h}^{-1})_j = \lceil n'/2 - 1 \rceil\right] \tag{5.9}$$

where Equation 5.7 holds because by definition the distribution of $\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,q,\tau}$ is exactly the same as $\mathcal{G}_\varepsilon^q$ under the condition $\|\underline{\zeta}\|_\infty \le \tau$.

Observe that, by the definition of $\underline{h}^{-1}$ and n', we may rewrite Equation 5.9 to obtain:

$$\Pr_{\underline{\zeta}}[B] \le \sum_{k=1}^{q} \Pr_{\underline{\zeta}}\left[\sum_{j=1}^{k} (\underline{h}(\underline{t}^{-1})_j + \zeta_j) = -b + \sum_{j=k+1}^{q} (\underline{h}(\underline{t}^{-1})_j + \zeta_j) + (q-2k)\tau\right] \tag{5.10}$$

where $b = 1$ if n' is even and $b = 0$ if n' is odd. Let us consider $b = 0$ (the other case follows by the same argument). By construction there are n players at position 0, so it follows that $h_1 = n$ and $h_j = 0$ for all $j > 1$. Therefore the RHS of Equation 5.10 equals

$$\Pr_{\underline{\zeta}}\left[n + \sum_{j=1}^{k} \zeta_j = \sum_{j=k+1}^{q} \zeta_j + (q-2k)\tau\right]$$

Since $\mathcal{G}_\varepsilon$ is symmetric therefore ζ_j is distributed identically as $-\zeta_j$, and combined with the above we may deduce:

$$\Pr_{\underline{\zeta}}[B] \le \sum_{k=1}^{q} \Pr_{\underline{\zeta}}\left[\sum_{j=1}^{q} \zeta_j = n - (q-2k)\tau\right] \tag{5.11}$$

We will prove the following lemma.

Lemma 5.5. *For all $k \in [q]$ and sufficiently large n, it holds that*

$$\Pr_{\underline{\zeta}}\left[\sum_{j=1}^{q} \zeta_j = n - (q-2k)\tau\right] \le e^{-1.9(1-\alpha)^2 n}$$

Applying this lemma to Equation 5.11 we obtain $\Pr_{\underline{\zeta}}[B] \le q e^{-1.9(1-\alpha)^2 n}$, which, for large n, is bounded by $e^{-(1-\alpha)^2 n}$. ■

We now turn to proving Lemma 5.5 The key point is the characterization of the sum of q two-sided geometric random variables given in Fact 3.3: $\sum_{j=1}^{q} \zeta_j$ is distributed identically to $Y - Y'$ where Y, Y' are independent $\mathcal{NB}_{q,\varepsilon}$ variables.

of Lemma 5.5. We will prove the following slightly stronger inequality:

$$\Pr_{\underline{\zeta}}\left[\sum_{j=1}^{q} \zeta_j \geq n-(q-2k)\tau\right] \leq e^{-1.9(1-\alpha)^2 n}$$

By Fact 3.3 it holds that $\sum_{j=1}^{q} \zeta_j$ is distributed identically to $Y - Y'$ as stated in Fact 3.3. Furthermore, since both Y, Y' are non-negative, $Y - Y' \geq n-(q-2k)\tau$ implies that $Y \geq n-(q-2k)\tau$. Therefore it suffices to bound $\Pr[Y \geq n-(q-2k)\tau]$. Furthermore, it suffices to consider just the case $k = 0$, which is the worst possible.

To summarize, it suffices to bound the probability $\Pr[Y \geq n - q\tau]$ where Y is a $\mathcal{NB}_{q,\varepsilon}$ variable. We apply the second point of fact Fact 3.3, which says that this probability is equal to the probability $\Pr[Z \leq q]$ where Z is a binomial random variable with $n - q\tau + q$ trials and success probability $1 - e^{-\varepsilon} = 1 - \alpha$. We can apply the Hoeffding bound for binomial variables and the fact that $q = \sqrt{n}$ and $\tau = o(\sqrt{n})$ (which follows from our hypothesis that $\eta = 2^{-o(\sqrt{n})/\log n}$) to conclude that, for sufficiently large n:

$$\Pr[Z \leq q] \leq e^{-2((1-\alpha)(n-q\tau)-\alpha q)^2/(n-q\tau+q)} \leq e^{-1.9(1-\alpha)^2 n} \tag{5.12}$$

■

Releasing the histogram leaks information Recall that M' discretizes the input into q intervals, samples $\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}$, computes the perturbed histogram $\underline{h}' = \underline{h}(\underline{t}) + \underline{\zeta} + \tau \cdot \underline{1}$, and outputs the median point of $\underline{h}'$.

We consider a slight modification of this mechanism: in addition to outputting the facility location, it also outputs the perturbed histogram $\underline{h}'$. Call this modified mechanism $\hat{M}$, and notice that $\hat{M}$ remains PTE: privacy holds because the perturbed histogram is $(2\varepsilon, \eta)$-differentially private, while truthfulness and efficiency remain (with respect just to the game value, before taking into account the information cost) because the facility location output is the same as what M' would have output. As mentioned in the introduction, we believe this represents a common practice: the database curator gathers the data using one particular incentive for the individuals, but in addition to the outcome of the game, he publishes some auxiliary information about the individuals' types that may be useful in the future for some other goal not necessarily related to the incentives used for the original database.

Theorem 5.6. *$\forall \varepsilon, \eta > 0$, suppose $\hat{M}$ is run with $(2\varepsilon, \eta)$-differential privacy. Then there exists $\eta' \leq \eta$ such that for any $\mathsf{IC}^{\eta'}_{\hat{M}}$ satisfying Assumption 5.1, for all inputs $\underline{t} \in Q^n$ and all $i \in [n]$, $\mathsf{IC}^{\eta'}_{\hat{M}}(\mathsf{Id}, \underline{t}, i) = \Omega(\varepsilon)$.*

Proof. $\hat{M}$ outputs a histogram perturbed by $\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}$. Given any database $\underline{t}$, construct the histogram $\underline{h}(\underline{t})$. The following Lemma 5.7 says that for the setting $\eta' = \Pr[\|\underline{\zeta}\|_\infty > \tau] \leq \eta$, for any player i, there exists $t' \neq t_i$ and B such that

$$\varepsilon < \log \frac{\Pr[\hat{M}(\underline{t}) \in B] - \eta'}{\Pr[\hat{M}(\underline{t}^{-i}, t') \in B]}$$

Theorem 5.6 then follows since Assumption 5.1 states that the information cost is at least the RHS (up to constant factors). ■

Lemma 5.7. *Fix any q a positive integer, $\underline{h} \in \mathbb{Z}^q$, $i \in [q]$. Let e_k denote the k'th standard basis vector. Suppose $\underline{\zeta}$ is sampled according to the distribution $\mathcal{H}_{\varepsilon,\tau,q}$, where τ satisfies $\Pr_{\underline{\zeta} \leftarrow \mathcal{G}^q_\varepsilon}[\|\underline{\zeta}\|_\infty > \tau] = \eta'$. Then for all $j \in [q]$, $j \neq i$, there exists $B \subseteq \mathbb{Z}^q$ such that:*

$$\log \frac{\Pr[\underline{h} + e_i + \underline{\zeta} \in B] - \eta'}{\Pr[\underline{h} + e_j + \underline{\zeta} \in B]} > \varepsilon$$

Proof. Let $B = \{\underline{h}' \mid h'_i > h_i\}$. Letting $\alpha = e^{-\varepsilon}$, we calculate that:

$$\begin{aligned}
\Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}}[\underline{h} + e_i + \underline{\zeta} \in B] &= \Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}}[\underline{\zeta}_i \geq 0] && (5.13)\\
&= \Pr_{\underline{\zeta} \leftarrow \mathcal{G}_\varepsilon^q}[\|\underline{\zeta}\|_\infty > \tau] + \Pr_{\underline{\zeta}_i \leftarrow \mathcal{G}_\varepsilon}[\tau \geq \underline{\zeta}_i \geq 0] && (5.14)\\
&= \eta' + \left(\tfrac{1-\alpha}{1+\alpha}\right) \sum_{j=0}^{\tau} \alpha^j && (5.15)\\
&> \eta' + \left(\tfrac{1-\alpha}{1+\alpha}\right) e^{\varepsilon} \sum_{j=1}^{\tau} \alpha^j && (5.16)
\end{aligned}$$

Above, Equation 5.14 holds because by the definition of $\mathcal{H}_{\varepsilon,\tau,q}$ (see Section 3), $\underline{\zeta}_i \geq 0$ can occur one of two ways: either we sampled $\underline{\zeta}' \leftarrow \mathcal{G}_\varepsilon^q$ and got $\|\underline{\zeta}'\|_\infty > \tau$ so we set $\underline{\zeta} = 0$, or else we sampled $\underline{\zeta}' \leftarrow \mathcal{G}_\varepsilon^q$ and got $\underline{\zeta}'_i \geq 0$ and we set $\underline{\zeta} = \underline{\zeta}'$. Similarly, we can deduce that:

$$\Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}}[\underline{h} + e_j + \underline{\zeta} \in B] = \Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}}[\underline{\zeta}_i \geq 1] < \left(\tfrac{1-\alpha}{1+\alpha}\right) \sum_{j=1}^{\tau} \alpha^j$$

Therefore, we may conclude that

$$\frac{\Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}}[\underline{h} + e_i + \underline{\zeta} \in B] - \eta'}{\Pr_{\underline{\zeta} \leftarrow \mathcal{H}_{\varepsilon,\tau,q}}[\underline{h} + e_j + \underline{\zeta} \in B]} > e^{\varepsilon}$$

■

Combining Theorem 5.3 and Theorem 5.6, we conclude that the following gives a deviation showing that $\hat{M}$ is *not* truthful when one uses the tradeoff utility of Equation 5.2 (for reasonable settings of ν_i).

Corollary 5.8 (Formalization of Theorem 2.5)**.** *Fix $\varepsilon > 0$ and $\eta = 2^{-o(\sqrt{n}/\log n)}$, and let $\hat{M}$ be the $(2\varepsilon, \eta)$-differentially private mechanism described above. There exists $\eta' \leq \eta$ such that the following holds for any $\mathsf{IC}^{\eta'}_{\hat{M}}$ satisfying Assumption 5.1 up to η approximation: if there exists a player i such that $\nu_i = \omega(e^{-(1-e^{-\varepsilon})^2 n})$. Let σ_0 be the strategy that always outputs 0. Then, there exists $i \in [n], \underline{t}^{-i} \in [0,1]^{n-1}$ such that for all $t_i \in [0,1]$, $\underline{t} = (\underline{t}^{-i}, t_i)$ satisfies $u_i(\mathsf{Id}, \underline{t}) < u_i(\sigma_0, \underline{t})$.*

Proof. Fix i such that $\nu_i = \omega(e^{-(1-\varepsilon)^2 n})$. Let $\underline{t}^{-i}$ be the input guaranteed to exist by Theorem 5.3. Using the definition of u_i (Equation 5.2) and applying Theorem 5.3, we have that for all $t_i \in [0,1]$:

$$\begin{aligned}
u_i(\mathsf{Id}, \underline{t}) &< \mathbb{E}_{(s,\underline{h}') \leftarrow \hat{M}(\underline{t}^{-i},0)}[v(t_i, s)] + e^{-(1-e^{-\varepsilon})^2 n} - \nu_i \varepsilon\\
&< \mathbb{E}_{(s,\underline{h}') \leftarrow \hat{M}(\underline{t}^{-i},0)}[v(t_i, s)] - \nu_i \cdot \mathsf{IC}_{\hat{M}}(\sigma_0, \underline{t}, i)\\
&= u_i(\sigma_0, \underline{t})
\end{aligned}$$

where we used the fact that $e^{-(1-e^{-\varepsilon})^2 n} - \nu_i \varepsilon < 0 = \mathsf{IC}_{\hat{M}}(\sigma_0, \underline{t}, i)$ (using Assumption 5.1). ■

This proves that, under the hypotheses of Corollary 5.8, not only is there $\underline{t}^{-i}$ such that player i would prefer not to tell the truth on some possible values of t_i (which we may view as "weakly" untruthful), but there is $\underline{t}^{-i}$ such that player i would *always* prefer to lie about his input for all values of t_i (which we may view as "strongly" untruthful).

5.2 Releasing synopses

Synopsis generators give a summary of the database that is accurate with respect to a specific set of count queries. Let $\mathcal{C}$ be a class of predicates on Q, *i.e.* functions mapping $Q \rightarrow \{0,1\}$. For $F \in \mathcal{C}, \underline{t} \in Q^n$, we define $\overline{F}(\underline{t}) = \frac{1}{n} \sum_{i=1}^{n} F(t_i)$.

Definition 5.9. M is a (γ, ρ)-*synopsis generator* on n-player inputs with respect to a class $\mathcal{C}$ if there is a real-valued function $P(s, F)$ such that, for all $\underline{t} \in Q^n$, $\Pr_{(s,p)\leftarrow M(\underline{t})}[\max_{F\in\mathcal{C}} |\overline{F}(\underline{t}) - P(s, F)| \leq \gamma] \geq 1 - \rho$.

Blum et al. [3] showed that, if $n = \tilde{O}(\frac{d \log |Q|}{\varepsilon\gamma^3})$ where d denotes the VC-dimension of $\mathcal{C}$, then there exists M that is a (γ, ρ)-synopsis generator with respect to $\mathcal{C}$ and also ε-differentially private (subsequent work [13, 21] improved the parameters).

5.2.1 Synopsis generators reveal information

Intuitively, it makes sense that if the synopsis must be accurate for a very rich class of predicates, then it must also be that the synopsis reveals a lot of information about the database. This is what we formalize next, by using the VC-dimension as a quantification of the “richness” of the class of predicates.

Definition 5.10. The *VC-dimension* of a class of predicates $\mathcal{C}$ is the largest d such that there exist $X = \{t_1, \ldots, t_d\} \subseteq Q$ and $D = 2^d$ functions $F_1, \ldots, F_D$ such that the F_i classify X in all possible ways, *i.e.* for all $b_1, \ldots, b_d \in \{0, 1\}$, there exists $j \in D$ such that $F_j(t_i) = b_i$. We say that X is the *shattering set* for $\mathcal{C}$, or that $\mathcal{C}$ shatters X.

Theorem 5.11. *Fix any $\gamma \in (0, \frac{1}{5}), \rho \in (0, 1)$. Suppose M is a mechanism that is a (γ, ρ)-synopsis generator on n-player inputs with respect to $\mathcal{C}$, which has VC dimension d. Suppose that IC_M satisfies Assumption 5.1. Then for all $\underline{t} \in Q^n$, there exists $\underline{t}' \in Q^n$ and $i \in [n]$ such that $\underline{t}, \underline{t}'$ differ in at most $4\gamma n$ entries and such that $\mathsf{IC}_M(\mathsf{Id}, \underline{t}', i) \geq \min(\Omega(\frac{d}{\gamma n}), \Omega(1))$.*

The intuition is the following: define the ball of radius γ induced by $\underline{t}$ in the outcome space, $B_\gamma(\underline{t}) = \{s \in S \mid \|P(s) - \overline{\mathcal{C}}(\underline{t})\|_\infty \leq \gamma\}$. We use a combinatorial design to construct a set of databases $T \subseteq Q^n$ of size $|T| \geq 2^{\Omega(d')}$ where $d' = \min(d, \gamma n)$, such that each $\underline{t}' \in T$ differ from $\underline{t}$ in at most $4\gamma n$ coordinates, and $B_\gamma(\underline{t}') \cap B_\gamma(\underline{t}'') = \varnothing$ for all $\underline{t}', \underline{t}'' \in T$. By the pigeon-hole principle, there exists $\underline{t}' \in T$, such that $\Pr[M(\underline{t}) \in B_\gamma(\underline{t}')] \leq 2^{-\Omega(d')}$. Since $\Pr[M(\underline{t}') \in B_\gamma(\underline{t}')] \geq 1 - \rho$ because M is a (γ, ρ)-synopsis generator, one of the hybrid databases between $\underline{t}, \underline{t}'$ must exhibit a large jump in the probability of giving an outcome in $B_\gamma(\underline{t}')$, and this gives a large information cost.

This proof is related to the packing arguments found in [22, 20, 6]; however, it is incomparable, as it gives quantitatively weaker bounds but it gives the additional property that, given *any* database, we can produce a *nearby* database that leaks a large amount of information. This last property was not present in previous lower bounds and is essential in the following applications.

of Theorem 5.11. By the definition of VC dimension, there exists a shattering set of size d. To simplify notation, let us name the shattering set $[d]$. The definition of VC dimension implies that for every $X \subseteq [d]$, there exists $F_X \in \mathcal{C}$ such that $F_X(t) = 1$ if $t \in X$ and $F_X(t) = 0$ if $t \in [d] \setminus X$. F_X can behave arbitrarily outside $[d]$.

Let $P(s) = (P(s, F))_{F\in\mathcal{C}}$ be the vector containing all estimates of counts. Let $\overline{\mathcal{C}}(\underline{t}) = (\overline{F}(\underline{t}))_{F\in\mathcal{C}}$. Let $B_\gamma(\underline{t}) = \{s \in S \mid \|P(s) - \overline{\mathcal{C}}(\underline{t})\|_\infty \leq \gamma\}$ be the γ-ball induced by $\underline{t}$ in the output space of the mechanism.

Lemma 5.12. *There exists an absolute constant $K > 0$ such that for any $1/5 > \gamma > 0$, and for all $\underline{t} \in Q^n$, there exists a set $T \subseteq Q^n$ satisfying:*

1. *$|T| \geq 2^{d'/K}$ where $d' = \min(d, 12\gamma n)$.*
2. *For all $\underline{t}' \in T$, there are exactly $4\gamma n$ coordinates i such that $t'_i \neq t_i$.*
3. *For all $\underline{t}', \underline{t}'' \in T$, it holds that $B_\gamma(\underline{t}') \cap B_\gamma(\underline{t}'') = \varnothing$.*

We first assume the lemma is true to prove the theorem. Let T be a set as guaranteed by Lemma 5.12. By the first and third properties, there exists $\underline{t}'' \in T$ such that

$$\Pr[M(\underline{t}) \in B_\gamma(\underline{t}'')] \leq 2^{-d'/K} \tag{5.17}$$

Fix such a $\underline{t}''$.

Let Z denote the set of $4\gamma n$ coordinates on which $\underline{t}$ and $\underline{t}''$ differ. Now consider the hybrids $\underline{t}^{(0)}, \dots, \underline{t}^{(4\gamma n)}$ where $\underline{t}^{(i)}$ agrees with $\underline{t}''$ on the first i coordinates in Z, and agrees with $\underline{t}$ on the last $4\gamma n - i$ coordinates in Z (and it agrees with both on the coordinates outsize Z). Clearly $\underline{t}^{(0)} = \underline{t}$ and $\underline{t}^{(4\gamma n)} = \underline{t}''$.

Let $\mathsf{wt}(i) = -\log \Pr[M(\underline{t}^{(i)}) \in B_\gamma(\underline{t}'')]$. We know that $\mathsf{wt}(0) \geq d'/K$ by Equation 5.17, and we know that $\mathsf{wt}(4\gamma n) \leq \log \frac{1}{1-\rho} \leq O(1)$ because M is a (γ, ρ)-synopsis generator, $\underline{t}'' = \underline{t}^{(4\gamma n)}$, and we assume that ρ is constant. Therefore we have

$$\tfrac{d'}{K} - O(1) \leq \mathsf{wt}(0) - \mathsf{wt}(4\gamma n) = \sum_{i=1}^{4\gamma n} \mathsf{wt}(i-1) - \mathsf{wt}(i)$$

By an averaging argument, there must exist i such that

$$\mathsf{wt}(i-1) - \mathsf{wt}(i) \geq \tfrac{1}{4\gamma n}\left(\tfrac{d'}{K} - O(1)\right) = \Omega\left(\tfrac{d'}{\gamma n}\right) = \min\left(\Omega\left(\tfrac{d}{\gamma n}\right), \Omega(1)\right)$$

Furthermore, by the definition of wt and Assumption 5.1 about the information cost IC_M, it therefore holds that $\mathsf{IC}_M(\mathsf{Id}, \underline{t}^{(i)}, j_i) \geq \Omega(\mathsf{wt}(i-1) - \mathsf{wt}(i))$, where j_i on the LHS equals the i'th element of Z and is the only coordinate that differs between $\underline{t}^{(i)}, \underline{t}^{(i-1)}$. Therefore, we deduce that $\mathsf{IC}_M(\mathsf{Id}, \underline{t}^{(i)}, j_i) \geq \min(\Omega(\frac{d}{\gamma n}), \Omega(1))$.

■

of Lemma 5.12. Let $\underline{h}(\underline{t})$ be the histogram of $\underline{t}$. Let us assume that $h_1 \leq h_2 \leq \ldots \leq h_d$ (for notational convenience and without loss of generality, since the names of the coordinates are immaterial and one could just rearrange them to satisfy this property).

Let $d'' \leq d$ be the largest integer such that $\sum_{i=1}^{d''} h_i \leq (1 - 4\gamma)n$. Either $d'' = d$, or else $d'' < d$ and we can deduce that:

$$n \geq \sum_{i=1}^{d} h_i = \sum_{i=1}^{d''+1} h_i + \sum_{i=d''+2}^{d} h_i \tag{5.18}$$

$$> (1-4\gamma)n + (d - d'' - 1)\frac{(1-4\gamma)n}{d''+1} \tag{5.19}$$

$$\Rightarrow \quad d'' \geq (1-4\gamma)d \tag{5.20}$$

Here we used the definition of d'' and the fact that the h_i are non-decreasing, meaning that h_i for $d \geq i > d'' + 1$ must satisfy $h_i \geq (1-4\gamma)n/(d''+1)$.

Set $d' = \min(d'', 12\gamma n)$. We first construct U which is a combinatorial design over $[d']$. Namely, $|U| \geq 2^{\Omega(d')}$ and each pair $X, Y \in U$ have small intersection.

1. Initially $U = \varnothing$, so pick an arbitrary $X \subseteq [d']$ of size $d'/3$. Add X to U.
2. If there exists $X \subseteq [d']$ such that $|X \cap Y| < d'/6$ for all $Y \in U$, then add X to U, otherwise halt and output U.

It is clear from the construction that, for all $X, Y \in U$, it holds that $|X \cap Y| < d'/6$. We show that U is exponentially large:

$$|U| \geq e^{2/18^2 \cdot d'/3} \tag{5.21}$$

This is a consequence of the Hoeffding inequality. Suppose we have already added i elements to U. We show that if $i < e^{2/18^2 \cdot d'/3}$ then there exists another subset that can be added. We use the probabilistic method by showing that the probability that a random subset X of $[d']$ with size $d'/3$ does not satisfy the desired property is strictly smaller than 1.

$$\Pr_{X \leftarrow \binom{[d']}{d'/3}}[\exists Y \in U, |X \cap Y| \geq d'/6] < e^{2/18^2 \cdot (d'/3)} \Pr[|X \cap Y| \geq d'/6] < 1$$

where we use the Hoeffding inequality (the version for sampling without replacement) in the final inequality (*i.e.* sampling $d'/3$ elements without replacement from among $[d']$, where the elements in Y are marked 1 and the rest are marked 0).

Let K be the smallest constant so that $|U| \geq 2^{d'/K}$. We now construct T using U. Let $X \in U$, then define $\underline{t}_X \in Q^n$ as follows. Let $X_1, \dots, X_{d'/3} \in [d']$ be the elements of X, say sorted in increasing order.

1. Initialize $i = 1$ and $j = 1$. (i will take value between $1, \dots, n$ and j between $1, \dots d'/3$).
2. Initialize $\underline{t}_X = \underline{t}$.
3. Do the following while $j \leq d'/3$:
    (a) Take the first $12\gamma n/d'$ individuals after and including the i'th individual in $\underline{t}_X$ whose values lie in $Q \setminus [d']$, and change their values to X_j.
    (b) Set i to be the individual after the last individual modified in the previous step, and increment j.

Observe two facts: first, we never "run out" of individuals to modify, since our choice of $d' \leq d''$ and Equation 5.20 ensure that the number of players with value in $Q \setminus [d']$ is at least $4\gamma n$. Second, $\frac{12\gamma n}{d'} \geq 1$ so in each iteration we modify at least one player.

We prove that $T = \{\underline{t}_X \mid X \in U\}$ satisfies the properties of the lemma. First, it is clear that if $X \neq Y$ then $\underline{t}_X \neq \underline{t}_Y$, and therefore $|T| \geq 2^{d/K}$. The second property holds because we modify $12\gamma n/d'$ individuals in each iteration, and there are $d'/3$ iterations.

To prove the third property, let $\underline{t}_X, \underline{t}_Y \in T$ be two distinct elements of T. Let $F_X \in \mathcal{C}$ be a function satisfying $F_X(x) = 1$ if $x \in X$ and $F_X(x) = 0$ if $x \in [d'] \setminus X$ (and F_X can behave arbitrarily outside $[d']$). Such F_X exists because $X \subseteq [d'] \subseteq [d]$ and $[d]$ is shattered by $\mathcal{C}$. Let $Z \subseteq [n]$ be the first $4\gamma n$ coordinates of $\underline{t}$ taking value in $Q \setminus [d']$. Observe that $\underline{t}$ and $\underline{t}_X$ are identical on all coordinates outside of Z. It holds that:

$$\begin{aligned}
|\overline{F}_X(\underline{t}_X) - \overline{F}_X(\underline{t}_Y)| &= \frac{1}{n}\left|\sum_{i=1}^{n}(F_X((\underline{t}_X)_i) - F_X((\underline{t}_Y)_i))\right| && (5.22)\\
&= \frac{1}{n}\left|\sum_{i\in Z} F_X((\underline{t}_X)_i) - \sum_{i\in Z} F_X((\underline{t}_Y)_i)\right| && (5.23)\\
&= \tfrac{1}{n}|\,|Z| - \tfrac{12\gamma n}{d'} \cdot |X \cap Y|\,| && (5.24)\\
&> \tfrac{1}{n}(4\gamma n - \tfrac{12\gamma n}{d'} \cdot \tfrac{d'}{6}) && (5.25)\\
&= 2\gamma && (5.26)
\end{aligned}$$

Suppose now for the sake of contradiction that $\exists s \in B_\gamma(\underline{t}_X) \cap B_\gamma(\underline{t}_Y)$. This means that $\|\overline{C}(\underline{t}_X) - P(s)\|_\infty \leq \gamma$ and $\|\overline{C}(\underline{t}_Y) - P(s)\|_\infty \leq \gamma$. But by the triangle inequality, this would imply that:

$$\begin{aligned}
2\gamma &< |\overline{F}_X(\underline{t}_X) - \overline{F}_X(\underline{t}_Y)|\\
&\leq |\overline{F}_X(\underline{t}_X) - P(s)| + |P(s) - \overline{F}_X(\underline{t}_Y)|\\
&\leq 2\gamma
\end{aligned}$$

which is a contradiction, and therefore $B_\gamma(\underline{t}_X) \cap B_\gamma(\underline{t}_Y) = \varnothing$. ■

5.2.2 Non-reactive mechanisms

As in Section 5.1, we would like to use Theorem 5.11 to infer that if the database owner publishes a synopsis rather than just the outcome of the game (in the hopes that the synopsis may be useful for other purposes), then individuals may prefer to lie because their gain in information cost outweighs their loss in value derived from the outcome. Intuitively this happens if by deviating, a player cannot lose too much value. We now formalize this.

Definition 5.13. M is (β, γ)-*non-reactive* if there exists $\underline{t} \in Q^n$ such that for all $\underline{t}'$ that differ from $\underline{t}$ in at most γn coordinates, for every $i \in [n]$ and $t'' \in Q$, it holds that $\mathbb{E}[v(t'_i, M(\underline{t}'))] \leq \mathbb{E}[v(t'_i, M((\underline{t}')^{-i}, t''))] + \beta$.

The following is an easy corollary of Theorem 5.11.

Corollary 5.14. *Fix any $\gamma \in (0, \frac{1}{5}), \rho \in (0, 1)$. If M is a (γ, ρ)-synopsis generator for $\mathcal{C}$ of VC-dimension d. Let $\nu = \min_i \nu_i$ and suppose that M is also $(o(\frac{\nu d'}{\gamma n}), 4\gamma)$-non-reactive, where $d' = \min(d, \gamma n)$. Then there exists $\underline{t} \in Q^n, i \in [n]$ and a strategy $\sigma(t)$ that is independent of t such that $u_i(\mathsf{Id}, \underline{t}) < u_i(\sigma, \underline{t})$.*

Proof. By the definition of non-reactive, let $\underline{t} \in Q^n$ be such that for all $\underline{t}'$ differing from $\underline{t}$ in at most $4\gamma n$ coordinates, for all $i \in [n], t'' \in Q$, it holds that

$$\mathbb{E}[v(t'_i, M(\underline{t}'))] \leq \mathbb{E}[v(t'_i, M((\underline{t}')^{-i}, t''))] + o(\tfrac{\nu d}{\gamma n})$$

By Theorem 5.11, one of these $\underline{t}'$ satisfies $\mathsf{IC}(\mathsf{Id}, \underline{t}', i) \geq \Omega(\frac{d}{\gamma n})$. Therefore, if we let σ be the strategy that outputs an arbitrary constant value in $x \in Q$ (and therefore by Assumption 5.1 it holds that $\mathsf{IC}_M(\sigma, \underline{t}', i) = 0$), we may write:

$$\begin{aligned} u_i(\mathsf{Id}, \underline{t}') &\leq \mathbb{E}[v(t'_i, M(\underline{t}'))] - \nu \cdot \mathsf{IC}(\mathsf{Id}, \underline{t}', i) \\ &< \mathbb{E}[v(t'_i, M((\underline{t}')^{-i}, x))] + o(\tfrac{\nu d}{\gamma n}) - \Omega(\tfrac{\nu d}{\gamma n}) \\ &< \mathbb{E}[v(t'_i, M((\underline{t}')^{-i}, x))] - \nu \cdot \mathsf{IC}(\sigma, \underline{t}', i) \\ &= u_i(\sigma, \underline{t}') \end{aligned}$$

Therefore, M is not truthful. ■

In particular, Corollary 5.14 holds even if M is differentially private as long as the VC-dimension of $\mathcal{C}$ is large. Blum et al. [3] prove that it is possible for M to be ε-differentially private and still be a (γ, ρ)-synopsis generator for a class of predicates with VC-dimension $d = \Omega(\frac{\gamma^3 \varepsilon n}{\log |Q| \log(1/\rho)})$. If in addition $|Q| = \mathrm{poly}(n)$ and $\nu = \Omega(1/\log |Q|)$ (which by Remark 5.2 constitutes a realistic setting of parameters), and M is $(o(1/\log n), 4\gamma)$-non-reactive, then M cannot be truthful.

5.2.3 The 1-facility location game on arbitrary bounded metric spaces

In fact, mechanisms may be quite non-reactive because intuitively the influence of a single individual on the outcome may diminish rapidly as there are more players. As a concrete example of such a class of mechanisms, we show that any efficient mechanism for a 1-facility location game over an arbitrary bounded metric space must be non-reactive. Let (Q, d) be a general bounded metric space, normalized so that

$$\max_{t,t' \in Q} \mathsf{d}(t, t') \leq 1$$

The general 1-facility location game is defined similarly to LINE-1-FAC, except that the type space is Q rather than just $[0, 1]$.

Theorem 5.15. *Suppose M is a δ-efficient mechanism for the 1-facility location game over a bounded metric space (Q, d). Then for any $\gamma \in (0, \frac{1}{2})$, it holds that M is $(\frac{2\delta}{(1-2\gamma)n-2}, \gamma)$-non-reactive.*

Therefore, we can construct the following deviation showing that M cannot be truthful if it is efficient for the 1-facility location on a metric space game and is also a good synopsis generator.

Corollary 5.16. *Let $\gamma \in (0, \frac{1}{10}), \rho \in (0, 1)$. Suppose M is a δ-efficient mechanism for the 1-facility location game on a bounded metric space, and also M is a (γ, ρ)-synopsis generator for $\mathcal{C}$ of VC-dimension d. Suppose that IC_M satisfies Assumption 5.1.*

Let $\nu = \min_i \nu_i$. If $\delta = o(\nu d')$ where $d' = \min(d, \gamma n)$, then there exists $\underline{t} \in Q^n, i \in [n]$ and a strategy $\sigma(t)$ that is independent of t such that $u_i(\mathsf{Id}, \underline{t}) < u_i(\sigma, \underline{t})$.

For example, the above theorem applies for the choice of parameters $d = \Omega(\frac{\gamma^3 \varepsilon n}{\log |Q| \log(1/\rho)})$, $\delta = n^{0.99}$, $|Q| = \mathrm{poly}(n)$, and $\nu = \Omega(1/\log |Q|)$.

Corollary 5.16 applies to a much broader setting than Corollary 5.8 in terms of the games considered. However, even when applied to LINE-1-FAC, Corollary 5.16 gives an incomparable result. Whereas Corollary 5.8 applies to the specific mechanism (Algorithm 4.10) studied in this paper, Theorem 5.15 holds for *any* efficient mechanism. On the other hand, Corollary 5.8 is better quantitatively, and also applies when only a histogram of the discretization of the player types is released, which may contain less information than a synopsis. (One can reconstruct the histogram from a synopsis if $\mathcal{C}$ is sufficiently rich, see for example Theorem 4.1 of [12].)

of Theorem 5.15. Let $\underline{t} \in Q^n$ be the vector where all coordinates have value x for an arbitrary $x \in Q$. Fix $\underline{t}' \in Q^n$ different from $\underline{t}$ in at most $m < n/2$ coordinates, *i.e.* m coordinates of $\underline{t}'$ are not equal to x. Suppose for convenience of notation that these are the first coordinates $t_1, \ldots, t_m$. We know that, by picking $s = x$, it is possible to achieve welfare $w(\underline{t}', x) \geq -\sum_{i=1}^{m} \mathsf{d}(x, t'_i)$, and therefore by the δ-efficiency of the mechanism, it holds that

$$\begin{aligned}
\delta &\geq w(\underline{t}', x) - \mathbb{E}_{s \leftarrow M(\underline{t}')}[w(\underline{t}', s)] && (5.27)\\
&\geq -\sum_{i=1}^{m} \mathsf{d}(x, t'_i) + \mathbb{E}_{s \leftarrow M(\underline{t}')}\left[(n-m) \cdot \mathsf{d}(s, x) + \sum_{i=1}^{m} \mathsf{d}(s, t'_i)\right] && (5.28)\\
&= \mathbb{E}_{s \leftarrow M(\underline{t}')}\left[(n-2m) \cdot \mathsf{d}(s, x)\right] + \mathbb{E}_{s \leftarrow M(\underline{t}')}\left[\sum_{i=1}^{m} (\mathsf{d}(s, t'_i) + \mathsf{d}(s, x) - \mathsf{d}(x, t'_i))\right] && (5.29)\\
&\geq \mathbb{E}_{s \leftarrow M(\underline{t}')}\left[(n-2m) \cdot \mathsf{d}(s, x)\right] \quad \text{(triangle inequality)} && (5.30)\\
&\Rightarrow \mathbb{E}_{s \leftarrow M(\underline{t}')}[\mathsf{d}(s, x)] \leq \frac{\delta}{n-2m} && (5.31)
\end{aligned}$$

For any $y \in Q$, we may write the following, using the triangle inequality and Equation 5.31:

$$\begin{aligned}
\mathbb{E}_{s \leftarrow M(\underline{t}')}[v(y, s)] &= -\mathbb{E}_{s \leftarrow M(\underline{t}')}[\mathsf{d}(y, s)] && (5.32)\\
&\leq -\mathbb{E}_{s \leftarrow M(\underline{t}')}[\mathsf{d}(y, x) - \mathsf{d}(s, x)] && (5.33)\\
&= -\mathsf{d}(y, x) + \mathbb{E}_{s \leftarrow M(\underline{t}')}[\mathsf{d}(s, x)] && (5.34)\\
&\leq v(y, x) + \tfrac{\delta}{n-2m} && (5.35)\\
\mathbb{E}_{s \leftarrow M(\underline{t}')}[v(y, s)] &= -\mathbb{E}_{s \leftarrow M(\underline{t}')}[\mathsf{d}(y, s)] && (5.36)\\
&\geq -\mathbb{E}_{s \leftarrow M(\underline{t}')}[\mathsf{d}(y, x) + \mathsf{d}(s, x)] && (5.37)\\
&= -\mathsf{d}(y, x) - \mathbb{E}_{s \leftarrow M(\underline{t}')}[\mathsf{d}(s, x)] && (5.38)\\
&\geq v(y, x) - \tfrac{\delta}{n-2m} && (5.39)
\end{aligned}$$

If $\underline{t}'$ differs from $\underline{t}$ in at most γn coordinates, then for every i, it holds that $((\underline{t}')^{-i}, t'')$ differs from $\underline{t}$ in at most $\gamma n + 1$ coordinates. Therefore we can apply Equation 5.35 and Equation 5.39 for $m = \gamma n + 1$ to obtain:

$$\begin{aligned}
\mathbb{E}[v(t'_i, M(\underline{t}'))] &\leq v(t'_i, x) + \tfrac{\delta}{(1-2\gamma)n-2}\\
&\leq \mathbb{E}[v(t'_i, M((\underline{t}')^{-i}, t''))] + \tfrac{2\delta}{(1-2\gamma)n-2}
\end{aligned}$$

■

6 Conclusions

In this paper we argued that the study of privacy must be coupled with the study of the incentives that motivate people to reveal private data. To study this question, we introduced a model combining differential privacy with truthfulness and efficiency. We constructed a general transformation that takes a truthful and efficienct mechanism and builds a PTE mechanism. We showed that when privacy is given an explicit numerical cost, even PTE mechanisms (which at first may seem to give us everything we want) are not necessarily sufficient. Our work has already spurred several follow-up works [4, 30, 23], and there are many interesting remaining open questions. For example, is it always possible for each problem to create a PTE mechanism that has as good efficiency as a mechanism that is not private? Another question, more empirical, is to explore what are good models for information cost functions and how we may be able to use a better understanding of these functions to achieve better parameters.

7 Acknowledgments

Part of this work was done while the author was visiting Microsoft Research Asia. The author would like to thank Pinyan Lu and Yajun Wang for their collaboration at the initial stages of this work. The author also thanks Salil Vadhan, Aaron Roth, and Yiling Chen for useful discussions, and the anonymous reviewers for their comments. The author was supported by the French ANR Defis program under contract ANR-08-EMER-012.

References

[1] Noga Alon, Michal Feldman, Ariel D. Procaccia, and Moshe Tennenholtz. Strategyproof approximation of the minimax on networks. *Mathematics of Operations Research*, 35(3):513–526, 2010.

[2] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank Mcsherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proc. of 26th PODS*, pages 273–282, 2007.

[3] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proc. 40'th STOC*, pages 609–618, 2008.

[4] 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.

[5] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 5:429–444, 1977.

[6] Anindya De. Lower bounds in differential privacy. In *TCC*, pages 321–338, 2012.

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

[8] Cynthia Dwork. Differential privacy. In *In Proc. ICALP*, pages 1–12. Springer, 2006.

[9] Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer Berlin / Heidelberg, 2008.

[10] Cynthia Dwork, Krishnaram Kenthapadi, Frank Mcsherry, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *In EUROCRYPT*, pages 486–503. Springer, 2006.

[11] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *In Proc. of the 3rd TCC*, pages 265–284. Springer, 2006.

[12] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. On the complexity of differentially private data release: Efficient algorithms and hardness results. In *Proc. 41'st STOC*, STOC '09, pages 381–390. ACM, 2009.

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

[14] Joan Feigenbaum, Aaron D. Jaggard, and Michael Schapira. Approximate privacy: foundations and quantification (extended abstract). In *Proc. 11th EC*, EC '10, pages 167–178, New York, NY, USA, 2010. ACM.

[15] Dan Feldman, Amos Fiat, Haim Kaplan, and Kobbi Nissim. Private coresets. In *Proc. 41st STOC*, STOC '09, pages 361–370, New York, NY, USA, 2009. ACM.

[16] I. Fellegi. On the question of statistical confidentiality. *J. of the Amer. Stat. Assoc.*, 67:7–18, 1972.

[17] Lisa K. Fleischer and Yu-Han Lyu. Approximately optimal auctions for selling privacy when costs are correlated with data. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 568–585, New York, NY, USA, 2012. ACM.

[18] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proc. 12th EC*, EC '11, pages 199–208, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0261-6.

[19] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proc. 41'st STOC*, STOC '09, pages 351–360. ACM, 2009.

[20] Moritz Hardt. *A Study of Privacy and Fairness in Sensitive Data Analysis*. PhD thesis, Princeton University, 2011.

[21] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proc. 51'st FOCS*, pages 61–70, Washington, DC, USA, 2010. IEEE Computer Society. doi: 10.1109/FOCS.2010.85.

[22] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proc. 42nd STOC*, pages 705–714, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0050-6. doi: 10.1145/1806689.1806786.

[23] Zhiyi Huang and Sampath Kannan. The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *Proc. FOCS '12*, 2012. To appear. Available at http://arxiv.org/abs/1204.1255.

[24] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *Proc. 49th FOCS*, pages 531–540, Washington, DC, USA, 2008. IEEE.

[25] Katrina Ligett and Aaron Roth. Take it or leave it: Running a survey when privacy comes at a cost. *CoRR*, abs/1202.4741, 2012.

[26] Pinyan Lu, Xiaorui Sun, Yajun Wang, and Zeyuan Allen Zhu. Asymptotically optimal strategy-proof mechanisms for two-facility games. In *Proceedings of the 11th ACM conference on Electronic commerce*, EC '10, pages 315–324. ACM, 2010.

[27] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual Symposium on Foundations of Computer Science*. Citeseer, 2007.

[28] H. Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35:437–455, 1980.

[29] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proc. 39th STOC*, pages 75–84, 2007.

[30] Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 774–789, New York, NY, USA, 2012. ACM.

[31] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. Approximately optimal mechanism design via differential privacy. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 203–213, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1115-1.

[32] Aaron Roth and Grant Schoenebeck. Conducting truthful surveys, cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 826–843, New York, NY, USA, 2012. ACM.

[33] J. Schummer and R. V. Vohra. Mechanism design without money. In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*, chapter 10, pages 243–266. Cambridge University Press, 2007.

[34] M. R. Spiegel. *Theory and Problems of Probability and Statistics*. McGraw-Hill, 1992.

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

```
Input: histogram $\underline{h}$ with $m$ coordinates. Let $n = \sum_{i=1}^{m} h_i$.
  1. Select a random permutation $\pi : [n] \to [n]$.
  2. For $i = 1$ to $n$, set $\hat{t}_{\pi(i)} = \min\{j \mid \sum_{k=1}^{j} h_k \geq i\}$. Let $\hat{\underline{t}}$ denote this setting of $\hat{t}_1, \ldots, \hat{t}_n$.
  3. Run $M(\hat{t}_1, \ldots, \hat{t}_n)$.
```

Algorithm A.1. Converting an arbitrary mechanism to one looking only at histogram.

A Mechanisms only need to consider the histogram

Suppose that G is an anonymous game, *i.e.* the global utility function is symmetric and all players have the same private utility functions. Suppose M is an arbitrary mechanism, possibly looking at individual types. We transform it into a mechanism M' that considers only the histogram according to Algorithm A.1.

M' is efficient because for all π it holds that the histogram of $\hat{\underline{t}}$ is exactly $\underline{h}$. Since the outcome of M is efficient on $\hat{\underline{t}}$, therefore the outcome is efficient for $\underline{h}$.

To see that M' is truthful, suppose that there were an input $\underline{h}$ with a possible deviation. Namely, there exists $\underline{t} \in Q^n, i \in [n], t'_i \in [n]$ such that $\mathbb{E}[u(t_i, M'(\underline{h}(\underline{t}^{-i}, t'_i)))] > \mathbb{E}[u(t_i, M'(\underline{h}(\underline{t})))]$. Let $\underline{h} = \underline{h}(\underline{t})$ and let $\underline{h}' = \underline{h}(\underline{t}^{-i}, t'_i)$. We will show using a (subtle) averaging argument that this implies that there is a deviation that allows some player to improve his utility for the original mechanism M.

Since we need only consider the case $t'_i \neq t_i$, we may define the following two quantities:

$$c = \sum_{k=1}^{t_i} h_k$$

$$c' = \sum_{k=1}^{t'_i} h_k$$

Namely, c is the number of players with type at most t_i (assuming the types are ordered $1, \ldots, q$), and likewise for c' with respect to t'_i.

Define the following cyclic permutation over n elements σ:

$$\sigma = \begin{cases} (c, c-1, \ldots, c'+1) & \text{if } c' < c \\ (c, c+1, \ldots, c') & \text{if } c' > c \end{cases}$$

Let M'_π be the mechanism of Algorithm A.1 with fixed permutation π. Then the assumption that $\underline{t}, i, t'_i$ constitutes a deviation for M' is equivalent to saying:

$$\mathbb{E}_{\pi \leftarrow S_n}[u(t_i, M'_{\pi \circ \sigma}(\underline{h}'))] > \mathbb{E}_{\pi \leftarrow S_n}[u(t_i, M'_\pi(\underline{h}))] \tag{A.1}$$

where S_n is the symmetric group over n elements.

By an averaging argument, there must exist π such that

$$\mathbb{E}[u(t_i, M'_{\pi \circ \sigma}(\underline{h}'))] > \mathbb{E}[u(t_i, M'_\pi(\underline{h}))] \tag{A.2}$$

where the expectation is only over the random coins used to run an execution of the original mechanism M.

We claim that this gives us a deviation for the original mechanism M. Let $\hat{t}_1 \ldots \hat{t}_n$ be the types that M' constructs using histogram $\underline{h}$ and permutation π, and let $\hat{\hat{t}}_1, \ldots, \hat{\hat{t}}_n$ be the types that M' constructs using histogram $\underline{h}'$ and permutation $\pi \circ \sigma$. Let $\ell = \pi(c)$.

Claim A.2. *For all $j \in [n]$, $j \neq \ell$ it holds that $\hat{t}_j = \hat{\hat{t}}_j$. Also it holds that $\hat{t}_\ell = t_i$ and $\hat{\hat{t}}_\ell = t'_i$.*

First observe that this gives a deviation for the original mechanism M, since plugging into Equation A.2 we get that

$$\mathbb{E}[u(\hat{t}_\ell, M(\underline{\hat{t}}^{-\ell}, \hat{\hat{t}}_\ell))] > \mathbb{E}[u(\hat{t}_\ell, M(\underline{\hat{t}}))]$$

where the randomness is only over the coins of M.

We now prove Claim A.2. Let us first consider the case that $c' < c$. We divide the analysis into the following cases:

1. For all $j \in \{\pi(1), \pi(2), \ldots, \pi(c')\}$: Fix any $z \in \{1, 2, \ldots, c'\}$. By the definition of σ and our assumption that $c' < c$, it holds that $z = \sigma(z)$. Therefore, since the $\underline{h}, \underline{h}'$ have the same counts for each type strictly smaller than t'_i, and $\underline{h}'_{t'_i} = \underline{h}_{t'_i} + 1$, it follows from the construction that $\hat{t}_{\pi(z)} = \hat{\hat{t}}_{\pi \circ (z)} = \hat{\hat{t}}_{\pi(z)}$.
2. For all $j \in \{\pi(c'+1), \pi(c'+2), \ldots, \pi(c-1)\}$: Fix any $z \in \{c'+1, c'+2, \ldots, c-1\}$. By the definition of σ and our assumption that $c' < c$, it holds that $z = \sigma(z+1)$. Therefore, since the $\underline{h}, \underline{h}'$ have the same counts for each type strictly smaller than t'_i, since $\underline{h}'_{t'_i} = \underline{h}_{t'_i} + 1$, and since $\underline{h}, \underline{h}'$ have the same counts on all types strictly between t'_i and t_i, it follows from the construction that $\hat{t}_{\pi(z)} = \hat{\hat{t}}_{\pi \circ \sigma(z+1)} = \hat{\hat{t}}_{\pi(z)}$.
3. For all $j \in \{\pi(c+1), \pi(c+2), \ldots, \pi(n)\}$: Fix any $z \in \{c+1, c+2, \ldots, n\}$. By the definition of σ and our assumption that $c' < c$, it holds that $z = \sigma(z)$. Therefore, since $\underline{h}, \underline{h}'$ have the same total number of elements whose type is strictly smaller than $c+1$, and since for each type between $c+1, n$ the two histograms have the same counts, it follows from the construction that $\hat{t}_{\pi(z)} = \hat{\hat{(t)}}_{\pi \circ \sigma(z)} = \hat{\hat{(t)}}_{\pi(z)}$.
4. For $\ell = \pi(c)$: By the definition of c, it holds that $\hat{t}_{\pi(c)} = t_i$. Observe that $\underline{h}, \underline{h}'$ have identical counts for all types strictly smaller than t'_i and $\underline{h}'_{t'_i} = \underline{h}_{t'_i} + 1$. By the definition of σ and our assumption that $c' < c$, it holds that $\sigma(c'+1) = c$. Therefore, it holds that $\hat{\hat{t}}_{\pi(c)} = \hat{\hat{t}}_{\pi \circ \sigma(c'+1)} = \hat{t}_{\pi(c')} = t'_i$.

We can repeat essentially the same analysis for the case that $c' > c$. (Observe that this case is not exactly symmetric with the previous case; in particular, in this case σ is a cycle of length $|c - c'| + 1$, while in the previous case it was of length $|c - c'|$. This is because of an asymmetry in our construction: the types $\hat{t}_j$ are created reading the histogram from left to right.) We divide the analysis into the following cases:

1. For all $j \in \{\pi(1), \pi(2), \ldots, \pi(c-1)\}$: Fix any $z \in \{1, \ldots, c-1\}$. By the definition of σ and our assumption that $c < c'$, it holds that $z = \sigma(z)$. Therefore, since the $\underline{h}, \underline{h}'$ have the same counts for each type strictly smaller than t_i, and $\underline{h}'_{t_i} = \underline{h}_{t_i} - 1$, it follows from the construction that $\hat{t}_{\pi(z)} = \hat{\hat{t}}_{\pi \circ (z)} = \hat{\hat{t}}_{\pi(z)}$.
2. For all $j \in \{\pi(c+1), \pi(c+2), \ldots, \pi(c')\}$: Fix any $z \in \{c+1, \ldots, c'\}$. By the definition of σ and our assumption that $c < c'$, it holds that $z = \sigma(z-1)$. Therefore, since the $\underline{h}, \underline{h}'$ have the same counts for each type strictly smaller than t_i, since $\underline{h}'_{t_i} = \underline{h}_{t_i} - 1$, and since $\underline{h}, \underline{h}'$ have the same counts on all types strictly between t_i and t'_i, it follows from the construction that $\hat{t}_{\pi(z)} = \hat{\hat{t}}_{\pi \circ \sigma(z-1)} = \hat{\hat{t}}_{\pi(z)}$.
3. For all $j \in \{\pi(c'+1), \pi(c'+2), \ldots, \pi(n)\}$: Fix any $z \in \{c'+1, c'+2, \ldots, n\}$. By the definition of σ and our assumption that $c < c'$, it holds that $z = \sigma(z)$. Therefore, since $\underline{h}, \underline{h}'$ have the same total number of elements whose type is strictly smaller than c', and since for each type between $c'+1, n$ the two histograms have the same counts, it follows from the construction that $\hat{t}_{\pi(z)} = \hat{\hat{(t)}}_{\pi \circ \sigma(z)} = \hat{\hat{(t)}}_{\pi(z)}$.
4. For $\ell = \pi(c)$: By the definition of c, it holds that $\hat{t}_{\pi(c)} = t_i$. By the definition of c' and σ and by our assumption that $c < c'$ (which is equivalent to saying $t_i < t'_i$), it holds that $\sigma(c') = c$ and furthermore the total count of elements of types at most t'_i is equal in $\underline{h}$ and $\underline{h}'$. Therefore it holds that $\pi\hat{\hat{(c)}} = \hat{\hat{t}}_{\pi \circ \sigma(c')} = \hat{t}_{\pi(c')} = t'_i$.

B Omitted Proofs

B.1 Proof of Lemma 3.2

of Lemma 3.2. From [11, 19], it holds for all $U \subseteq \mathbb{Z}^q$ that

$$\Pr[\underline{\zeta} \in U] \leq e^{2\varepsilon} \Pr[e_i - e_j + \underline{\zeta} \in U] \tag{B.1}$$

Our situation is almost the same, except our perturbed histogram has the following distribution: sample $\underline{\zeta} \leftarrow \mathcal{G}^q_\varepsilon$ and check whether $\|\underline{\zeta}\|_\infty > \tau$. If so, set $\underline{\zeta}' = 0$, otherwise set $\underline{\zeta}' = \underline{\zeta}$.

Let $X_a = \{\underline{x} \in \mathbb{Z}^q \setminus \{0\}, \|\underline{x}\|_\infty \leq a\}$, the set of all non-zero points with infinity norm at most a. We will use the observation that, by the definition of ζ', for all $\underline{x} \in X_\tau$, it holds that $\Pr[\underline{\zeta} = \underline{x}] = \Pr[\underline{\zeta}' = \underline{x}]$, and also $\Pr[\underline{\zeta} = 0] \leq \Pr[\underline{\zeta}' = 0]$.

Fix an arbitrary set U, and divide up U into three disjoint parts: $U_0 = U \cap \{0\}$, $U_1 = U \cap X_{\tau-1}$, and $U_2 = U \setminus U_0 \setminus U_1$. We reason about each of the three separately. Let $\alpha = e^{-\varepsilon}$.

1. By the definition of $\underline{\zeta}'$, $\Pr[\underline{\zeta}' = 0] = \Pr[\zeta = 0] + \Pr[\|\zeta\|_\infty > \tau]$. Therefore, Equation B.1 and the fact that $-e_i + e_j \in X_\tau$ imply that

   $$\begin{aligned}\Pr[\zeta = 0] &\leq e^{2\varepsilon} \Pr[e_i - e_j + \zeta = 0] \\ &= e^{2\varepsilon} \Pr[e_i - e_j + \zeta' = 0]\end{aligned}$$

   Along with Equation 3.1, this implies that

   $$\Pr[\underline{\zeta}' = 0] \leq e^{2\varepsilon} \Pr[e_i - e_j + \zeta' = 0] + \tfrac{2q\alpha^{\tau+1}}{1+\alpha}$$

2. Using Equation B.1, it holds that

   $$\begin{aligned}\Pr[\underline{\zeta}' \in U_1] &= \Pr[\underline{\zeta} \in U_1] && \text{(B.2)} \\ &\leq e^{2\varepsilon} \Pr[e_i - e_j + \underline{\zeta} \in U_1] && \text{(B.3)}\end{aligned}$$

   Since $U_1 \subseteq X_{\tau-1}$, it holds that the shifted set $U_1 - e_i + e_j$ is contained in the set $X_\tau \cup \{0\}$. Therefore it follows that

   $$\Pr[e_i - e_j + \underline{\zeta} \in U_1] \leq \Pr[e_i - e_j + \underline{\zeta}' \in U_1]$$

   which, combined with Equation B.3 implies

   $$\Pr[\underline{\zeta}' \in U_1] \leq e^{2\varepsilon} \Pr[e_i - e_j + \underline{\zeta}' \in U_1]$$

3. Since $\underline{\zeta}'$ takes range in $[-\tau, \tau]^q$, we have that

   $$\begin{aligned}\Pr[\underline{\zeta}' \in U_2] &= \Pr[\|\underline{\zeta}\|_\infty = \tau] \\ &\leq \frac{2q(1-\alpha)\alpha^\tau}{1+\alpha} \\ &\leq e^{2\varepsilon} \Pr[e_i - e_j + \underline{\zeta}' \in U_2] + \frac{2q(1-\alpha)\alpha^\tau}{1+\alpha}\end{aligned}$$

Combining all three sets gives us

$$\begin{aligned}\Pr[\underline{\zeta}' \in U] \leq{}& e^{2\varepsilon} \Pr[e_i - e_j + \underline{\zeta}' \in U] \\ &+ \frac{2q(1-\alpha)\alpha^\tau + 2q\alpha^{\tau+1}}{1+\alpha}\end{aligned}$$

which in turn implies the lemma. ■

B.2 Proof of Theorem 4.7

of Theorem 4.7. The mechanism is listed in Algorithm 4.6.

Truthfulness. Since the the players are symmetric, it suffices just to consider the truthfulness of player 1. Fix $\underline{t}^{-1} = (t_2, \ldots, t_n)$. Let $\underline{h} = \underline{h}(\underline{t})$.

For all j it holds that $h_j \geq 0$. Furthermore, because player 1 is in column t_1, it holds that $h_{t_1} \geq 1$.

The mechanism's output is the minimal $s \geq 1$ such that $\sum_{j=1}^{s} h_j \geq n/2$. Let s be the output of the mechanism, and we consider what happens when t_1 declares some other value t^*. Let $\underline{h}^*$ be the histogram that is identical to $\underline{h}$ everywhere, except $h^*_{t_1} = h_{t_1} - 1 \geq 0$ and $h^*_{t^*} = h_{t^*} + 1$. Let s^* be the minimal s such that $\sum_{j=1}^{s^*} h^*_j \geq n/2$, namely the output of the mechanism on input $(\underline{t}^{-1}, t^*)$. We analyze the following cases, using the fact that both $\underline{h}, \underline{h}^*$ are non-negative:

1. $t_1 < s$. Because for all $s' < s$ it holds that $\sum_{j=1}^{s'} h^*_j \leq \sum_{j=1}^{s-1} h_j < n/2$, it follows that $s^* \geq s$. Since $t_1 < s$, this implies that $v(t_1, s^*) \leq v(t_1, s)$. Furthermore, observe that if $s^* > s$ then $v(t_1, s^*) \leq v(t_1, s) - 1$.

2. $t_1 = s$: in this case player 1's utility is 0, which cannot be improved (since for this game the utility is a non-positive number). Furthermore, if $s^* \neq s$ then $v(t_1, s^*) \leq -1 \leq v(t_1, s) - 1$.

3. $t_1 > s$: Because it holds that $\sum_{j=1}^{s} h^*_j \geq \sum_{j=1}^{s} h_j \geq n/2$, therefore $s^* \leq s$. Since $t_1 > s$, this implies that $v(t_1, s^*) \leq v(t_1, s)$. Furthermore, if $s^* < s$ then $v(t_1, s^*) \leq v(t_1, s) - 1$.

Therefore, regardless of the value of t_1, it holds that $v(t_1, s^*) \leq v(t_1, s)$, and therefore player 1 has no incentive to misreport his type. Furthermore, if $s^* \neq s$ then $v(t_1, s^*) \leq v(t_1, s) - 1$.

Efficiency. Let s be the output of the mechanism. We prove the utility is greater for s than for all other s'.

Claim B.1. *For any histogram $\underline{h}$ and $s = M(\underline{h})$, and for all $s' \in [q]$, it holds that $\sum_{j=1}^{q} h_j |j - s'| \geq \sum_{j=1}^{q} h_j |j - s|$.*

Recalling that $v(t, s) = -\gamma |t - s|$, this immediately implies that $w(\underline{t}, s') \leq w(\underline{t}, s)$ for all s'.

We now prove the claim. First consider $s' < s$. Split the summation $\sum_{j=1}^{q} h_j |j - s'|$ into three parts:

$$\sum_{j=1}^{q} h_j |j - s'| = \sum_{j \leq s'} h_j (s' - j) + \sum_{s' < j < s} h_j (j - s') + \sum_{s \leq j \leq q} h_j (j - s') \tag{B.4}$$

We will bound each of the three terms. The following hold:

$$\sum_{j \leq s'} h_j (s' - j) = \sum_{j \leq s'} h_j (s - j) - (s - s') \sum_{j \leq s'} h_j \tag{B.5}$$

$$\sum_{s \leq j} h_j (j - s') = \sum_{s \leq j} h_j (j - s) - (s' - s) \sum_{s \leq j} h_j \tag{B.6}$$

$$\begin{aligned} \sum_{s' < j < s} h_j (j - s') &= \sum_{s' < j < s} h_j (s - j) - \sum_{s' < j < s} h_j (s + s' - 2j) \\ &\geq \sum_{s' < j < s} h_j (s - j) - (s - s') \sum_{s' < j < s} h_j \end{aligned} \tag{B.7}$$

The first two equalities follow by definition; Equation B.7 is justified by the inequality $s + s' - 2j < s - s'$ and the fact that the h_j are non-negative.

Applying Equation B.5 Equation B.6, Equation B.7 to Equation B.4, we obtain

$$\sum_{j=1}^{q} h_j |j - s'| \geq (s - s') \left(\sum_{s \leq j} h_j - \sum_{j < s} h_j \right) + \sum_{j=1}^{q} h_j |j - s| \tag{B.8}$$

Since $s - s' > 0$, $\sum_{j=1}^{q} h_j = n$, and, by the definition of s it holds that $\sum_{j<s} h_j < n/2$, it follows that the first term on the RHS of Equation B.8 is positive. This implies the claim for the case $s' \le s$.

The case for $s' \geq s$ follows similarly, resulting in the inequality

$$\sum_{j=1}^{q} h_j |j - s'| \geq (s' - s) \left(\sum_{s<j} h_j - \sum_{j \leq s} h_j \right) + \sum_{j=1}^{q} h_j |j - s|$$

Now, using the fact that $s' \geq s$ and by the definition of s, it holds that $\sum_{s<j} h_j \leq n/2$, we can similarly conclude that the claim also holds in this case. ■

C Further discussion about information cost definition

In [4, 30], a subtle point is raised about the definition of information cost. In our work, we assume that the information cost is zero if the player reports a type that is independent of his true type, and this requires that the information cost depend on the strategy used by the player. Intuitively, this models what an adversary, knowing the strategy the player used, would deduce about the players' private information after seeing the output of the mechanism. This based on the following rationality argument: if the adversary knows that the players will behave strategically, then in analyzing the adversary we should use a measure where the adversary assumes that the players played their optimal strategy when trying to learn private information from the output of the mechanism. In contrast, Chen et al. [4], Nissim et al. [30] argue that one should use a measure where the adversary tries to learn private information assuming that the players reported truthfully. We believe that both points of view have merit, and which view to take depends on whether, in a particular context, the adversary assumes that the players played strategically or not.

C.1 Max-divergence based candidate

We propose two possible measure of information cost. The first uses the following general information-theoretic measure, which is essentially a "max-divergence" measure between the output distributions of the mechanism on different player inputs (shifted slightly to be non-negative). The max-divergence was introduced in the context of differential privacy by Dwork et al. [13], and we refer the reader there for a discussion. We work with the η-approximate notion here in order to prevent the information cost from being artificially large (or even infinite) due to some low-probability events.

Definition C.1. The η-approximate max-divergence information cost of strategy σ to player i on input $\underline{t}$ and for mechanism M is denoted $\mathsf{IC}_M^{\eta}(\sigma, \underline{t}, i)$ and defined as

$$\max_{t' \in Q} \max_{\substack{B \subseteq S \times \mathbb{R}^n \\ \Pr[M(\underline{t}^{-i}, \sigma(t)) \in B] > \eta}} \log \frac{\Pr[M(\underline{t}^{-i}, \sigma(t)) \in B] - \eta}{\Pr[M(\underline{t}^{-i}, \sigma(t')) \in B](1 - \eta)}$$

We write IC_M to denote IC_M^0.

Observe that this definition satisfies Assumption 5.1. A simple calculation shows that if a mechanism is (ε, η)-differentially private, then it holds that $\mathsf{IC}_M(\sigma, \underline{t}, i) \leq \varepsilon + \log \frac{1}{1-\eta}$ for all $\sigma, \underline{t}, i$. Note that typically we will take η to be negligible and so we can consider $\log \frac{1}{1-\eta} \approx 0$.

C.2 Mutual information based candidate

If one puts a probability distribution T on player i's input, then one can measure the information cost using the following natural measure based on information cost, which appeared in previous versions of this paper:

Definition C.2. The mutual-information-based information cost of strategy σ to player i on input $\underline{t}$ with prior distribution on inputs T and with respect to mechanism M is:

$$\mathsf{IC}_M(\sigma, \underline{t}, i) = I(T; M(\underline{t}^{-i}, \sigma(T)))$$

The definition satisfies the first part of Assumption 5.1, namely that players can hide their data. However, this definition does not satisfy the second part, namely that cost reflects differential privacy (Equation 5.1). This is because this definition is inherently Bayesian, whereas Equation 5.1 requires some worst-case properties. Our impossibility results about releasing histograms do generalize to this measure (Corollary 5.8) although with a more involved proof, while as far as we know our impossibility results on synopsis generators (Corollary 5.16) do not extend to this measure.