Optimal Deviations for Strongly Privacy-Aware Agents: When Differential Privacy is Counterproductive

Esther Plotnick Asher Spector*

December 2019

Abstract

In the context of mechanism design, we model the utility functions of agents who value their privacy. Our utility function is motivated by agents who do not want the mechanism *facilitator* to know their private information but have strong incentives to participate in a mechanism. This contrasts with previous research, which focused on constructing differentially private mechanisms which do not leak agent's information to the outside world, but still give the mechanism facilitator access (or the appearance of access) to agent's reports.

We analyze agent's behavior under these utility functions in three ways. First, we characterize optimal strategies for agents which balance their incentives to participate in a mechanism against their privacy concerns. Second, we make some informal remarks on when these strategies constitute Bayes-Nash Equilibria. Lastly, we show via simulations that under these utility functions, differentially-private mechanisms incentivize agents to misreport *more* than non-differentially private mechanisms, leading to net lower social utility. This phenomenon occurs because individual reports are unlikely to change the output of differentially-private mechanisms, meaning that players are incentivized to lie to preserve their privacy, since it will not affect the outcome anyway.

1 Introduction

1.1 Motivation

A growing body of literature in mechanism design has acknowledged that agents often have preferences for privacy: in addition to preferring specific possible outcomes, agents may not want others to know their private information.

Previous literature has largely focused on the case where a realized outcome reveals private information about agents (see [Chen et al., 2011], [Huang and Kannan, 2012], [Pai and Roth, 2013], [Xiao, 2013], and [Sui and Boutilier, 2011]). Indeed, many standard mechanism design settings have potential privacy losses, such as when private valuations for items/packages in an auction are revealed through open(ed) bids. We refer to such individuals as **weakly privacy-aware**. Researchers have designed various mechanisms which incentivize weakly privacy-aware agents to report (approximately) truthfully to a mechanism, and then the mechanism outcome does not leak information about any individual players to the outside world ([Huang and Kannan, 2012] and [Xiao, 2013] for differentially private mechanisms).

However, in this class paper, we consider the case where individuals do not want the mechanism *facilitator* to know their private information at all – we refer to such individuals as **strongly**

*Author order determined alphabetically

privacy aware. For example, users answering a social media poll may not want Facebook to know their private information, even if Facebook promises to only use differentially private algorithms on the data (especially since users may not believe this promise). At the same time, users may have a countervailing incentive to participate in the poll in order to affect the outcome of the mechanism. To understand these tradeoffs, this class paper defines and analyzes a new model of strongly privacy-aware utility functions.[1]

We are certainly not experts on the mechanism design/privacy literature, but to our limited knowledge, very little previous research has analyzed this question.

1.2 Structure and Main Claims

This paper is one third survey paper and two thirds novel contribution. It focuses exclusively on settings without payments (we have no good reason for this – we just ran out of time to analyze payments). It is structured as follows:

- In section 2, we present a model of strongly privacy-aware utility functions. Unlike previous work, this model applies in cases where the agent prefers that the mechanism facilitator does not have access to their private information.
- In section 3, we present a brief and highly incomplete survey of differential privacy in mechanism design, both as a tool to create approximately-truthful mechanisms and to protect information. For a more comprehensive survey, see [Pai and Roth, 2013].
- In section 4, we characterize optimal strategies for privacy-aware agents given priors on other player's reports, and offer some remarks on whether/when these strategies form Bayes-Nash Equilibrium.
- In section 5, we simulate these deviations for the 1-dimensional facility location problem (which we will define later), and demonstrate that in many settings differentially-private mechanisms lead to both lower social utility and greater misreports.
- In section 6 (the conclusion), we acknowledge some limitations of our model/simulations and discuss possible ways to remedy them.

2 Model and Problem Statement

In this section, we'll (i) define some preliminary notation, (ii) state a model of agent's utility which incorporates their concerns about privacy, (iii) discuss why we think it's a reasonable model, and (iv) more precisely outline our goals for the rest of the class paper.

2.1 Preliminary Notation

Given a set of N agents labeled $i \in \{1, 2, \dots N\}$, we define the following notation:

- Let $\mathcal{O}$ be the set of *outcomes* ($o \in \mathcal{O}$).

[1] To be clear, we are not conceptualizing this as an information elicitation problem, which certainly been analyzed previously – we consider the case where agents face a genuine trade-off between caring about the outcome of a mechanism and preserving their privacy.

- Let $\mathcal{V}$ be the set of *types*, which we also assume to be the set of allowable reports. Let $|\mathcal{V}| = V$. Let the true type of player i be $v_i \in \mathcal{V}$. Let the report of a player (not necessarily truthful) be $\hat{v}_i \in \mathcal{V}$.
- Let $\Delta(\mathcal{V})$ be the set of *strategies* ($s \in \Delta(\mathcal{V})$).
- Let $v : \mathcal{O} \to \mathbb{R}^+$ be a *valuation function* with v_i that for agent i.
- Finally, for any strategy $s = (s_1, \ldots, s_V) \in \Delta(\mathcal{V})$, define the *Shannon Entropy* of s as

  $$H(s) = -\sum_{i=1}^{V} s_i \log(s_i)$$

  As we'll discuss later, the entropy of a distribution quantifies the amount of uncertainty or "randomness" in the distribution. For example, a degenerate distribution (e.g. a constant value) has entropy of $\log(1) = 0$, and a perfectly uniform distribution over V values has entropy $-\log(1/V) = \log(V)$, representing the "maximal" uncertainty for any distribution over V values.

  To this end, we will often work with the "normalized" entropy $H(s)/\log(V)$, which ranges between 0 and 1, with 1 corresponding to the maximal possible uncertainty (a discrete uniform distribution).
- Lastly, it's worth noting for the theory section that we can generalize the Shannon entropy to be between two distributions to the cross entropy

  $$H(s, s^*) = -\sum_{i=1}^{V} s_i^* \log(s_i)$$

  and the KL divergence

  $$KL(s||s^*) = H(s, s^*) - H(s)$$

  It is also well known that the KL divergence is minimized (optimizing over s) when $s = s^*$.

2.2 Utility Function Model

We define $u : \mathcal{O} \times \Delta(\mathcal{V}) \to \mathbb{R}^+$ be a *modified utility function* with u_i that for agent i, which combines outcome valuation with privacy valuation. We model *strongly* privacy-aware agents' utilities as:

$$u(o, s) = v(o) + \alpha \cdot \frac{H(s)}{\log V}$$

Note that for a fixed V, working with the normalized entropy makes no difference, since one could scale α to cancel it out: we just think it makes some of the plots later a bit more interpretable.

Interpretation: For an outcome o and strategy s, the modified utility $u(o, s)$ is the sum of the value of the outcome $v(o)$ (standard utility) and the normalized Shannon entropy of s. This utility function will usually force players to make trade-offs between their valuation functions and their privacy (e.g., a player who reports truthfully may favorably influence the outcome, but will reveal lots of private information). We can think of the scale parameter α as a measure of the agents' value for privacy: an agent who doesn't value their privacy might have $\alpha \approx 0$, and an agent who only cares about privacy might have a very large α.

Why might entropy be a good way to represent agent's concerns about privacy? We picked entropy for three reasons.

1. First, the entropy of a strategy quantifies an observer's uncertainty about the initial player's type, even if they know the initial player's report. This is in stark contrast to utility functions proposed by [Xiao, 2013], [Chen et al., 2011], which only model information leaked by the outcome of the mechanism, as opposed to the player's report.

   As a simple example, set $V = 4$ and consider two players with strategies $s_1 = (0.5, 0.5, 0, 0)$ and $s_2 = (0.25, 0.25, 0.25, 0.25)$, which have normalized entropies of 0.5 and 1. Even if an observer knew the players' reports, as opposed to just the mechanism output, they could only down player 1 and 2's types to one of 2 and 4 types respectively, independent of the mechanism. As a result, regardless of the mechanism, our model indicates that player 2 has twice as much "privacy" as player 1.

2. Second, this utility function is a nice extension to the differential privacy literature. As described in the background section, [Huang and Kannan, 2012] have shown that common differentially private mechanisms optimize a weighted sum of social utility and the entropy of the *outputs* of a mechanism. Our utility function represents a stronger notion of privacy, since players balance their own valuation functions against the entropy of their *reports.*

3. Third, this utility function is easy to analyze - as we'll see in section 4, it makes it tractable to characterize best responses for players in very general settings.

There are certainly some limitations to this model, but we will discuss them in the conclusion.

2.3 Precise Problem Statement

Given the above model of strongly privacy-aware utility functions, we are interested in answering the following three questions.

1. First, suppose player i has a prior $\pi_i \in \Delta(\mathcal{V})$ on the distribution of other agent's reports. What is the strategy s_i^* for player i which optimizes their expected privacy-aware utility, e.g. which appropriately balances their valuation function against their privacy?

2. Second, can we find approximate Bayes-Nash Equilibria for these utility functions?

3. Third, we are interested in the social choice function

$$g : \mathcal{V}^N \to \mathcal{O},\ g(\hat{v}) = \arg\max_{o \in \mathcal{O}} \sum_{i=1}^{N} \hat{v}_i(o)$$

   Let $\mathcal{M} : \mathcal{V}^N \to \mathcal{O}$ be a (direct revelation) *mechanism* that selects an outcome $o \in \mathcal{O}$ for reports $(\hat{v}_1, \ldots \hat{v}_N) = \hat{v} \in \mathcal{V}$. $\hat{v}_{-i} = (\hat{v}_1, \ldots \hat{v}_{i-1}, \hat{v}_{i+1}, \ldots \hat{v}_n)$ is standard.

   We will define a "differentially private" mechanism in the next section; however, for now, it's enough to know that the outcomes of differentially private mechanisms depend only very weakly on individual player's reports. We hypothesize that weakly private (e.g. differentially private) mechanisms incentivize privacy-aware players to *lie*, because their report is unlikely to effect the outcome, so they choose to preserve their privacy given expected low cost to lying.

We will give a theoretical treatment of question (1) and some remarks about (2) in section 4, and we give some preliminary empirical evidence in section 5 that our hypothesis is correct (e.g., that differentially private mechanisms may incentivize misreporting for privacy-concerned agents).

In the following section 3, we first give a brief survey of related work and introduce one mechanism from the literature in detail.

3 Differentially Private Mechanism Design

This class paper began as an exposition of literature in privacy and mechanism design, investigating optimal strategies and efficiency concerns that arise when agents act to protect their privacy. This section reflects this expositional goal, though the following work takes a different approach from our model in that it frames the problem in terms of *differential privacy*, and is motivated by weakly privacy-aware agents.

3.1 Differential Privacy

We use the framework of differential privacy mechanisms from survey [Pai and Roth, 2013] and we use the same notation as introduced previously. Additionally, let $\mathcal{R}$ be the probability space from which a mechanism takes its randomness, and let two types $t, t' \in \mathcal{V}^N$ be *neighbors* if there exists an agent i such that $t_{-i} = t'_{-i}$; either $t = t'$ or the types are equal except for the type of agent i.

Definition 3.1 (ϵ-Differential Private Mechanism). *A mechanism* $: \mathcal{V} \times \mathcal{R} \to \mathcal{O}$ *is* $\epsilon-$*differentially private*[2] $(0 < \epsilon \ll 1)$ *if for all neighboring types* $t, t' \in \mathcal{V}^N$ *and for all utility functions* $u : \mathcal{O} \to \mathbb{R}^+$

$$\mathbb{E}_{o \sim \mathcal{M}(t)}[u(o)] \leq exp(\epsilon)\mathbb{E}_{o \sim \mathcal{M}(t')}[u(o)]$$

Intuitively, this differential privacy guarantee implies that for any utility function the change of a report can only have a small $\approx 1 + \epsilon$ scaling effect on the expected utility of the outcome of mechanism $\mathcal{M}$.

Early work in this area aimed to use differential privacy as a *tool* for mechanism design. McSherry and Talwar showed that an ϵ-differentially private mechanism is also 2ϵ-approximately dominant strategy truthful, and they used this to build approximately dominant strategy truthful mechanisms [McSherry et al., 2007]. This connection is intuitive by definition; differential privacy guarantees that an individual input will have a small effect on the expected outcome, and so likewise misreporting could only have small utility gains for an agent. Unfortunately, the advantage of this mechanism property could be superficial, as there is then also little incentive to tell the truth; thus, agents with preferences for privacy may just choose to misreport to preserve their privacy.

Another useful property of differentially private mechanisms is resistance to collusion. We can extend the definition of a type's neighbor to consider types $t, t' \in \mathcal{V}$ with differences in k indices. Then we have that $\mathbb{E}_{o \sim \mathcal{M}(t)}[v(o)] \leq exp(k\epsilon)\mathbb{E}_{o \sim \mathcal{M}(t')}[v(o)]$. So, k changes in a type only change the expected outcome by $\approx 1 + k\epsilon$ for $k \ll 1/\epsilon$. This property applies to coalitions of k agents and gives us a mechanism that is resistant to collusion.

For any mechanism design setting that aims to maximize social welfare, [Huang and Kannan, 2012] demonstrates that one particular mechanism – the exponential mechanism (the details of which we do not discuss here) – can be instantiated as a truthful mechanism that also preserves differential privacy.

3.2 Differentially Private Mechanisms

We describe a differentially private mechanism, which we will later simulate. There are many other frameworks for differentially private mechanisms, which, given more time, we could also analyze.

[2] For our purposes this definition is most relevant, although it is not the standard definition for general differential privacy; [Pai and Roth, 2013] say that the definition is "easily seen to be equivalent."

3.2.1 Discrete Facility Location, [Chen et al., 2011]

The discrete facility location problem takes agents' location reports and chooses a location for a facility. We only consider choosing one facility in the one dimensional case for simplicity/tractability.

With similar notation as to our model's framework, let $\mathcal{L}$ be the set of locations with $|\mathcal{L}| = L$ and label $l_j \in \mathcal{L}$ for $j \in \{1, 2, \dots L\}$. For N agents $i \in \{1, 2, \dots N\}$, let $\ell \in \mathcal{L}^N$ be the true agents' locations, let $\hat{\ell} \in \mathcal{L}^N$ be the agents' reports, and let $f \in \mathcal{L}$ be the chosen facility location $Loc(\hat{\ell})$ (mechanism $Loc : \mathcal{L}^N \to \mathcal{L}$). We assume that agents' valuation functions are $v_i(\ell, f) = c - |f - \ell_i|$, where c is some constant designed to ensure that $v_i(\ell, f) \geq 0$ for any ℓ and f. For this context, it is well-established that the truthful and efficient mechanism for the discrete facility location problem is to the select the median of the reports.[3]

We now introduce the privacy framework from [Chen et al., 2011], which sets up a differentially private *truthful* mechanism for the discrete facility location problem. We frame truthfulness with the notation for the discrete facility location problem for simplicity.

Definition 3.2 (Truthful Mechanism)**.** *For all players $i \in \{1, \dots N\}$, all locations $\ell \in \mathcal{L}^N$ and all possible reports $\hat{\ell} \in \mathcal{L}^N$ (any $\hat{\ell}'_{-i} \in \mathcal{L}^{N-1}$ and $\hat{\ell}_i \in \mathcal{L}$), Loc is* truthful *for player i if*[4]

$$v_i(\ell, Loc(\ell_i, \hat{\ell}'_{-i})) \geq v_i(\ell, Loc(\hat{\ell}_i, \hat{\ell}'_{-i}))$$

An analogous definition exists for truthfulness in expectation for randomized mechanisms. Let *LocRand* now be a randomized mechanism $Loc : \mathcal{L}^n \times \mathcal{R} \to \mathcal{L}$ (recall that $\mathcal{R}$ is the probability space from which a mechanism would take its randomness).

Definition 3.3 (Truthful Mechanism in Expectation)**.** *For all players $i \in \{1, \dots N\}$, all locations $\ell \in \mathcal{L}^N$ and all possible reports $\hat{\ell} \in \mathcal{L}^N$ (any $\hat{\ell}'_{-i} \in \mathcal{L}^{N-1}$ and $\hat{\ell}_i \in \mathcal{L}$), LocRand $:^N \times \mathcal{R} \to \mathcal{L}$ is* truthful in expectation *for player i if*

$$\mathbb{E}[v_i(\ell, LocRand(\ell_i, \hat{\ell}'_{-i}))] \geq \mathbb{E}[v_i(\ell, LocRand(\hat{\ell}_i, \hat{\ell}'_{-i}))]$$

We say that *LocRand* is *universally truthful* if the inequality holds for all values $r \in \mathcal{R}$.

[Chen et al., 2011] gives a framework and examples of differentially private, truthful mechanisms. The following mechanisms is for the discrete facility location problem.

Mechanism. (Chen12 Mechanism)**.**

Input: privacy parameter ϵ, reports $\hat{\ell} \in \mathcal{L}^N$

1. First, take the frequencies of the reported types. [Chen et al., 2011] describe this as a histogram $h = (h_1, \dots h_L)$ where h_j is the frequency of reports of $l_j \in \mathcal{L}$.
2. Next, choose random noise $r \in \mathbb{N}^L$ such that $\Pr[r_j = k]$ is proportional to $exp(-\epsilon k/2)$. (Each component of r is independent).
3. Finally, consider the perturbed histogram $h + r$, which is just the originally frequencies with added noise.

[3] Note that some of the privacy concerns within the facility location problem extend beyond distance-based valuations and median allocation. Consider the setting in which the facility is a hospital that would specialize in a specific health concern. If an agent has this health concern, this agent would presumably have higher value for the facility but may also have greater incentive to protect this health information.

[4] Note that this definition is different from a Nash equilibrium, which considers the incentives of a player given that the other players use equilibrium strategies.

Output: $Median(h+r)$, the median of the noisy histogram Practically, taking the median consstitutes the following function:

$$Median(\hat{\ell}) = \min_{k \in [L]} \left\{ k : \sum_{j=1}^{k} \hat{\ell}_j \geq \sum_{j=k+1}^{L} \hat{\ell}_j \right\}$$

Theorem 3.1 ([Chen et al., 2011]). *The Chen12 mechanism is ϵ-differentially private.*

The Chen12 mechanism is also universally truthful and individually rational for some constraints on the privacy utility function and the outcome utility function ([Chen et al., 2011]). However, we will show below that it is not truthful for strongly privacy-aware agents.

4 Theoretical Results

We begin with some notation. Suppose a player has true valuation function v_i. Given a strategy $s_i \in \Delta(\mathcal{V})$, define the *expected value* to agent i given i.i.d. reports v_{-i} from the agent's prior π_i as

$$F_{\pi_i}(v_i, s_i) = \mathbb{E}_{v_i \sim s_i}\left(\mathbb{E}_{v_{-i} \sim \pi_i}\left[v_i(\mathcal{M}(v_i', v_{-i}))\right]\right)$$

Moreover, we abuse notation slightly and use v_i' to refer to the degenerate strategy which always reports v_i', so

$$F_{\pi_i}(v_i, v_i') = \mathbb{E}_{v_{-i} \sim \pi_i}\left[v_i(\mathcal{M}(v_i', v_{-i}))\right]$$

Now we can characterize optimal deviations for players. Note that the proof technique used in Theorem 4.1 is *not* novel (as [Huang and Kannan, 2012] have noted, it's a well known result in statistical physics), but we include it because it's very important for this paper.

Theorem 4.1 (Optimal Strategies for Privacy Aware Agents). *For a player with true valuation v_i, a prior π_i on other player's reports, and utility function $u(o,s) = v_i(o) + \alpha \cdot H(s)/\log(V)$ where $\alpha > 0$, define strategy $s^* \in \Delta(V)$ such that*

$$s_j^* \propto \exp\left(\frac{\log(V) \cdot F_{\pi_i}(v_i, v_j)}{\alpha}\right)$$

where s_j^* corresponds to the probability of reporting v_j. Then s^* ***uniquely*** *maximizes Player i's expected utility, e.g.*

$$s^* = \arg\max_s \mathbb{E}_{v_i' \sim s}\left[\mathbb{E}_{v_{-i} \sim \pi_i} u(\mathcal{M}(v_i', v_{-i}), s)\right] = \arg\max_s F_{\pi_i}(v_i, s) + \frac{\alpha \cdot H(s)}{\log(V)}$$

Note in particular that s^* follows a Gibbs-like distribution.

Proof. First we know definitionally that

$$\mathbb{E}_{v_i' \sim s}\left[\mathbb{E}_{v_{-i} \sim \pi_i} u(\mathcal{M}(v_i', v_{-i}), s)\right] = \sum_{v_j \in \mathcal{V}} F_{\pi_i}(v_i, v_j) \cdot s_j + \frac{\alpha}{\log(V)} \cdot H(s)$$

Playing with the left term, we obtain

$$= \frac{\alpha}{\log(V)} \sum_{v_j \in \mathcal{V}} s_j \cdot \log\left(\exp\left(F_{\pi_i}(v_i, v_j) \cdot \frac{\log(V)}{\alpha}\right)\right) + \frac{\alpha}{\log(V)} \cdot H(s)$$

This looks like the distribution s_j^* without a normalizing constant, so we multiply and divide by the normalizing constant inside the log. For ease of notation, define $r_j^* = \exp\left(\frac{F_{\pi_i}(v_i,v_j)\cdot\log(V)}{\alpha}\right)$ to be the unnormalized s_j^*, so $s_j^* = r_j^*/(\sum_{l=1}^V r_l^*)$. Then our expression becomes

$$= \frac{\alpha}{\log(V)} \sum_{v_j \in \mathcal{V}} \left[s_j \cdot \log\left(\frac{r_j^*}{\sum_{l=1}^V r_l^*}\right) + s_j \log\left(\sum_{l=1}^V r_l^*\right)\right] + \frac{\alpha}{\log(V)} H(s)$$

Note first that $r_j^*/\sum_{l=1}^V r_j^* = s_j^*$ by definition, so we can substitute that in. Second, the new normalizing term doesn't depend on j, so we can pull it out of the sum (since $\sum_{v_j \in \mathcal{V}} s_j = 1$ by definition):

$$= \frac{\alpha}{\log V}\left[\sum_{v_j \in \mathcal{V}} s_j \log\left(s_j^*\right) + H(s)\right] + \frac{\alpha}{\log V} \log\left(\sum_{l=1}^V r_l^*\right)$$

Note however that the very first sum on the left equal to the negative cross entropy between s^* and s (we defined it a while back in section 2)! This yields

$$= \frac{\alpha}{\log(V)}\left(-H(s,s^*) + H(s)\right) + \frac{\alpha}{\log V} \log\left(\sum_{l=1}^V r_l^*\right)$$

and plugging in the definition of the KL divergence,

$$= -\frac{\alpha}{\log(V)} KL(s||s^*) + \frac{\alpha}{\log(V)} \log\left(\sum_{l=1}^V r_l^*\right)$$

Note that we are interested in maximizing this quantity with respect to s. Since s does not appear in the right hand term, this is equivalent to maximizing

$$-\frac{\alpha}{\log(V)} KL(s||s^*)$$

which is equivalent to minimizing the KL divergence between s and s^*. The solution to this, as noted in section 1, is to set $s = s^*$, and moreover, this solution is unique. Therefore s^* uniquely maximizes the expected utility of the player. □

An immediate corollary is that for this class of utility functions, for *any* priors and bounded valuation function, no mechanism without payments is exactly truthful for $\alpha > 0$.

Corollary 4.1.1. *Suppose player i has an arbitrary valuation function $v_i : O \to \mathbb{R}^+$ and utility function $u(o,s) = v_i(o) + \alpha/\log(V) \cdot H(s)$ for $\alpha > 0$. Suppose also that $v_i(o) \leq M \in \mathbb{R}$ for some M. Then for any prior π_i and any mechanism $\mathcal{M}$, player i's best response is not truthful.*

Proof. Note that the best response s_j^* as defined above satisfies

$$s_j^* \propto \exp\left(F_{\pi_i}(v_i,v_j) \cdot \log(V)/\alpha\right) > 0$$

where the strict inequality holds because $F_{\pi_i}(v_i,v_j)$ must be bounded because v_i is bounded, and $\alpha > 0$. This holds for all j, meaning that player i's best strategy involves misreporting with some positive probability. □

This is a remarkably strong claim: many impossibility theorems (such as the Gibbard–Satterthwaite (GS) theorem) state that for certain classes of mechanisms (e.g. nondictatorial onto mechanisms with more than three outputs), there are *some* situations where there are useful deviations for players. In contrast, this corollary proves that in *every* situation, no matter their uncertainty/beliefs about other player's actions, (e.g. maybe π_i is degenerate, maybe it is not), privacy-aware agents should employ a randomized strategy.

Of course, as α approaches 0, this Gibbs distribution does rapidly approach a degenerate distribution and may be approximately truthful.

Remark 1. Is This an Equilibrium? Suppose players $1, \ldots, N$ have priors $\pi_1, \ldots, \pi_N$, and that they then create a set of strategies $(s_1^*, \ldots, s_N^*) \in \Delta(\mathcal{L})^N$.

It's worth noting that the set of strategies do not form a Bayes-Nash equilibrium in the conventional sense. If π_i represents player i's prior over the *true locations* of the other players, and player i assumes that the other players are not being truthful, then s_i^* will not usually be a best response, because s_i^* is only a best response when π_i is the distribution of player i's responses.

However, the strategies $(s_1^*, \ldots, s_N^*)$ are optimal if the priors $\{\pi_i\}$ represent priors on other player's reports as opposed to their true locations. There's a plausible argument that this makes more sense than treating $\{\pi_i\}$ as priors on the true locations, since players also likely have significant uncertainty about which strategies other agents will use; treating $\{\pi_i\}$ as priors on reports better incorporates this uncertainty.

Remark 2. Finding Traditional BNEs. Despite the last remark, it would be interesting to find traditional BNEs. One way to do this would be to find convenient priors π on the true locations of other players satisfying the following property: define $s^*(\ell_j, \pi) \in \Delta(\mathcal{L}) \subset \mathbb{R}^L$ as the best strategy for a player with true valuation v_j and prior $\pi \in \Delta(\mathcal{L}) \subset \mathbb{R}^L$, and suppose that

$$\sum_{j=1}^{L} \pi_j \cdot s^*(\ell_j, \pi) = \pi$$

In other words, π is equal to a sum of the vectors $s^*(\ell_j, \pi)$ weighted by the components π_j. If this equality holds, then the strategies $(s_1^*, \ldots, s_N^*)$ would form a conventional Bayes-Nash Equilibrium, because even if another player creates optimal misreports for each particular valuation function/type they could have, their *marginal* distribution of reports (averaged over their possible valuation functions) would still be π.

Unfortunately, we found it rather difficult to find priors π satisfying this property, since s^* has a complex dependency with π. However, we suspect that in simple cases (e.g. where the mechanism $\mathcal{M}$ is constrained to have some kind of symmetry), such priors exist, and often they are reasonable (e.g. close to uniform).

Conjecture 4.1.1. *For some classes of mechanisms, it is possible to find priors π on player's true locations such that $s_1^*, \ldots, s_N^*$ are a conventional BNE.*

Remark 3. Empirical Note. In practice, in the simulations below, we find that

$$\sum_{j=1}^{V} \pi_j \cdot s^*(v_j, \pi) \approx \pi$$

for a uniform π, meaning that the strategies $s_1^*, \ldots, s_N^*$ are an approximate BNE.

5 Comparison Simulations

We hypothesized initially that under this model of utility functions, differentially private algorithms would lead to *worse* social outcomes, because they rely *less* on individual reports and therefore give weaker incentives to individuals to be truthful. In this section, we confirm this hypothesis for the discrete facility location problem by running simulations comparing the vanilla median mechanism to the Chen 12 mechanism described in section 3.2.1.

If you want to follow along/rerun experiments, all of our code is available at `https://github.com/amspector100/c136privacyproject`, with instructions in the README.

5.1 Simulation Design

We simulate the 1-dimensional location facility problem as described in section 3.2.1. Given an arbitrary location mechanism $\mathcal{M}$, simulations are run as follows:

1. We start by picking L evenly spaced locations between -1 and 1.
2. We sample the true locations of N players independently (we take N to be small so players have some influence on the output). Order the possible locations as $l_{(1)} < \cdots < l_{(L)}$. We add some skew into this distribution, so for each player, the probability of having $l_{(i)}$ selected as their true location is proportional to i^3. We only do this once (we set the seed to 136 and then hold these true locations constant for each Monte Carlo sample in the upcoming steps).
3. We assume each player has a uniform prior $\pi_i = (1/L, \ldots, 1/L)$. Although this prior is not technically correct (the actual location distribution is skewed), we think that's fairly realistic: if people's priors always matched the true distribution, most statistical problems would be pretty trivial. Using these players' priors, we use Monte-Carlo sampling (incorporating the mechanism $\mathcal{M}$) to obtain precise estimates for $F_{\pi_i}(v_i, v_j)$ for each i and j.
4. For each player i, we find the optimal strategy $s_i^* \in \Delta(L)$ by setting $s_{ij}^* \propto \exp\left(\frac{\log(V) \cdot F_{\pi_i}(v_i, v_j)}{\alpha}\right)$, as in Theorem 4.1.
5. Finally, we use Monte-Carlo sampling again to calculate the expected total social utility (excluding utility from privacy), defined as follows:

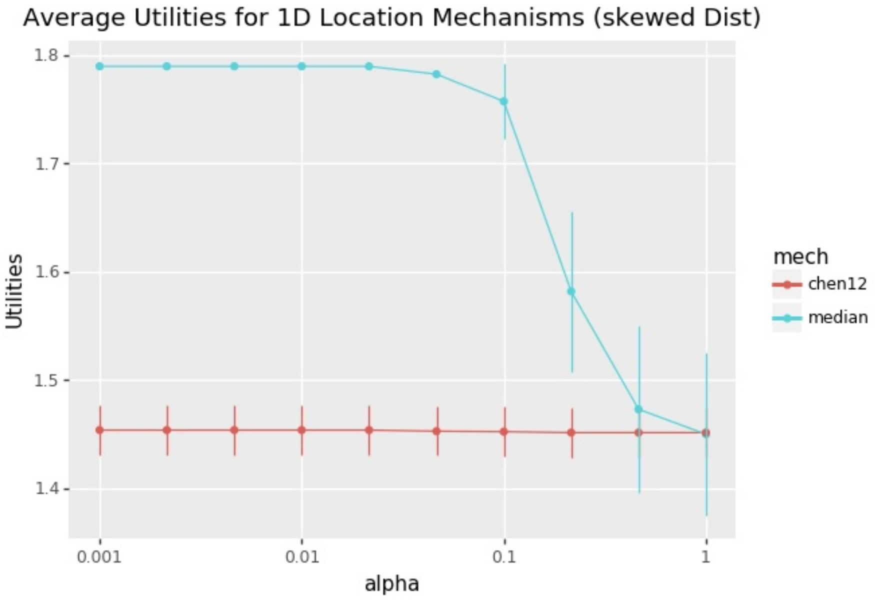
$$V(\mathcal{M}, \pi) = \mathbb{E}_{\hat{v}_i \sim s_i^*}\left[\sum_{i=1}^{N} v_i(\mathcal{M}(\hat{v}))\right]$$

   In particular, we sample independently from these optimal strategies, apply the mechanism $\mathcal{M}$ to obtain an output o, and calculate the average social utility $\frac{1}{N}\sum_{i=1}^{N} v_i(o)$, and then average over the samples.

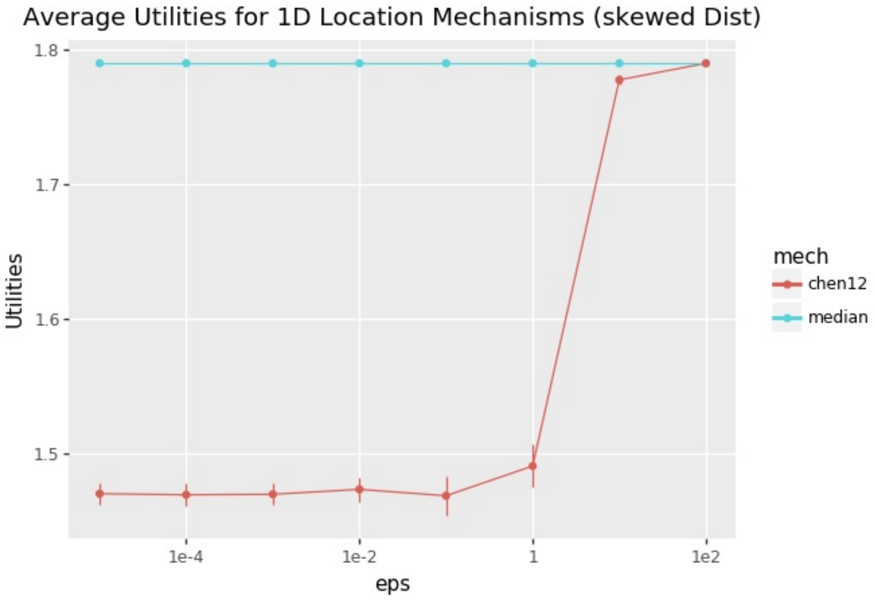
By default, we set the ϵ parameter of the Chen 12 mechanism to be 0.1, $L = 100$, and $\alpha = 0.1$, although we'll present plots varying each parameter. In steps 3 and 5, we use 100 Monte-Carlo samples (and we report standard errors).

5.2 Results

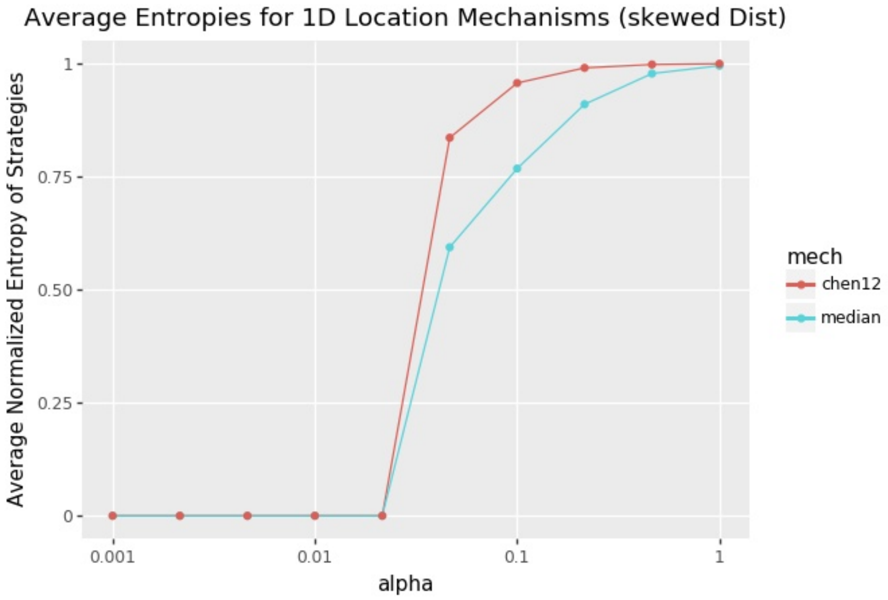
As expected, we find that the differentially private Chen12 mechanism leads to substantially lower social utility than the vanilla median mechanism. This is displayed in the figure below. Note that as described above, the "utilities" are equal to a constant (2) minus the distance from the selected mechanism to the player's location and *do not* account for privacy. The maximum utility is therefore 2 and the minimum utility is $2 - |1 - -1| = 0$.



As α gets higher and higher, note that player's reporting distributions converge to uniform (as they value their privacy more and more), and as a result, the median mechanism starts to look identical to the Chen12 mechanism. Note that the effect disappears as the Chen12 becomes less differentially private (and converges in probability to the median mechanism).



Although the efficiency loss of the Chen12 mechanism is to some extent inevitable (differential privacy usually comes at an efficiency cost), we also find that players are *less* truthful under the Chen12 mechanism, presumably because it relies less on individual player reports and therefore gives players less of an incentive to report truthfully. One way to see this is that both on average and for each player, the normalized entropy of the optimal strategy is higher in the Chen12 mechanism:



6 Conclusion and Future Work

This paper has three main branches of analysis. First, we modelled the utility functions of strongly privacy-aware agents concerned with reporting their private data regardless of differential privacy guarantees on the outcome (which they care enough about to participate in the mechanism). Second, we characterized optimal strategies for these agents in very general settings, and showed they follow a Gibbs-like distribution. Third and finally, we showed that often these strategies are *more* random (and therefore less truthful) when responding to differentially private mechanisms, because players' inputs have a less direct effect on the outcome, so they have less incentive to report truthfully.

We discuss some limitations of our work below:

- First, it is unclear whether our utility model leads to fully realistic on the behalf of players. For example, players who highly value their privacy may simply decide not to participate at all. One way to remedy this might be allow a new strategy, called ϕ, which represents not participating, and give this strategy a fixed utility.
- Second, in repeated games, strategies with high entropy may not preserve your privacy. For example, if a player has type 3 of three possible types, the strategy $(0.25, 0.25, 0.5)$ has higher entropy than the strategy $(0, 0.5, 0.5)$, but in a repeated game, only the latter strategy will full conceal the player's type. We would be interested in exploring Bayesian interpretations of this.
- There are also many avenues of further exploration, which include: (i) characterizing mechanisms and priors that give a BNE strategy (Conjecture 4.1.1) (ii) evaluating efficiency losses on other differentially private mechanisms (iii) accounting for privacy alongside efficiency to analyze privacy/efficiency tradeoffs.

Thanks again for a wonderful semester and for getting to page 12 of our class paper!

References

[Chen et al., 2011] Chen, Y., Chong, S., Kash, I. A., Moran, T., and Vadhan, S. P. (2011). Truthful mechanisms for agents that value privacy. *CoRR*, abs/1111.5472.

[Huang and Kannan, 2012] Huang, Z. and Kannan, S. (2012). The exponential mechanism for social welfare: Private, truthful, and nearly optimal. *CoRR*, abs/1204.1255.

[McSherry et al., 2007] McSherry, F., McSherry, F., and Talwar, K. (2007). Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA. IEEE Computer Society.

[Pai and Roth, 2013] Pai, M. M. and Roth, A. (2013). Privacy and mechanism design. *CoRR*, abs/1306.2083.

[Sui and Boutilier, 2011] Sui, X. and Boutilier, C. (2011). Efficiency and privacy tradeoffs in mechanism design. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, pages 738–744. AAAI Press.

[Xiao, 2013] Xiao, D. (2013). Is privacy compatible with truthfulness? In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 67–86, New York, NY, USA. ACM.