Privacy Games

Yiling Chen*, Or Sheffet**, and Salil Vadhan***

Center for Research on Computation and Society
School of Engineering and Applied Sciences
Harvard University
{yiling,osheffet,salil}@seas.harvard.edu

Abstract. The problem of analyzing the effect of privacy concerns on the behavior of selfish utility-maximizing agents has received much attention lately. Privacy concerns are often modeled by altering the utility functions of agents to consider also their privacy loss [24,13,19,4]. Such privacy aware agents prefer to take a randomized strategy even in very simple games in which non-privacy aware agents play pure strategies. In some cases, the behavior of privacy aware agents follows the framework of Randomized Response, a well-known mechanism that preserves differential privacy.

Our work is aimed at better understanding the behavior of agents in settings where their privacy concerns are explicitly given. We consider a toy setting where agent A, in an attempt to discover the secret type of agent B, offers B a gift that one type of B agent likes and the other type dislikes. As opposed to previous works, B's incentive to keep her type a secret isn't the result of "hardwiring" B's utility function to consider privacy, but rather takes the form of a payment between B and A. We investigate three different types of payment functions and analyze B's behavior in each of the resulting games. As we show, under some payments, B's behavior is very different than the behavior of agents with hardwired privacy concerns and might even be deterministic. Under a different payment we show that B's BNE strategy does fall into the framework of Randomized Response.

1 Introduction

In recent years, as the subject of privacy becomes an increasing concern, many works have discussed the potential privacy concerns of economic utility-maximizing agents. Obviously, utility-maximizing agents are worried about the effect of revealing personal information in the current game on future transactions, and wish to minimize potential future losses. In addition, some agents may simply care about what some outside observer, who takes no part in the current game, believes about them. Such agents would like to optimize the effect of their behavior in the current game on the beliefs of that outside observer. Yet specifying the exact way in which information might affect the agents' future payment or an outside observer's beliefs is a complicated and intricate task.

* Supported in part by NSF grant CCF-1301976.

** Supported in part by NSF grant CNS-1237235.

*** Supported by NSF grant CNS-1237235, a gift from Google, Inc., and a Simons Investigator grant.

Differential privacy (DP), a mathematical model for privacy, developed for statistical data analysis [9,8], avoids the need for such intricate modeling by providing a worst-case bound on an agents' exposure to privacy-loss. Specifically, by using a ϵ-differentially private mechanism, agents can guarantee that the belief of *any* observer about them changes by no more than a multiplicative factor of $e^{\epsilon} \approx 1 + \epsilon$ once this observer sees the outcome of the mechanism [7] . Furthermore, as pointed out in [13,19], using a ϵ-differentially private mechanism the agents guarantee that, in expectation, *any* future loss increases by no more than a factor of $e^{\epsilon} - 1 \approx \epsilon$. A recent line of work [24,13,19,4] has used ideas from differential privacy to model and analyze the behavior of privacy-awareness in game-theoretic settings. The aforementioned features of DP allow these works to bypass the need to model future transactions. Instead, they model privacy aware agents as selfish agents with utility functions that are "hardwired" to trade off between two components: a (positive) reward from the outcome of the mechanism vs a (negative) loss from their non-private exposure. This loss can be upper-bounded using DP, and hence in some cases can be shown to be dominated by the reward (of carefully designed mechanisms), showing that privacy concerns don't affect an agent's behavior.

However, in other cases, the behavior of privacy-aware agents may differ drastically from the behavior of classical, non-privacy aware agents. For example, consider a toy-game in which B tells A which of the two free gifts that A offers (or *coupons* as we call it, for reasons to be explained later) B would like to receive. We characterize B using one of two types, 0 or 1; where type 0 prefers the first gift and type 1 prefers the second one. (This is a rephrasing of the "Rye or Wholewheat" game discussed in [19].) Therefore it is simple to see that a non-privacy-aware agent always (deterministically) asks for the gift that matches her type. In contrast, if we model the privacy loss of a privacy-aware agent using DP as in the work of Ghosh and Roth [13] (and the value of the coupon is large enough), a privacy-aware agent takes a randomized strategy. (See Section 2.2.) Specifically, the agent plays *Randomized Response*, a standard differentially private mechanism that outputs a random choice slightly biased towards the agent's favorable action.

However, it was argued [19,4] that it is not realistic to use the worst-case model of DP to quantify the agent's privacy loss and predict her behavior. Differential privacy should only serve as an *upper bound* on the privacy loss, whereas the agent's expected privacy loss can (and should in fact) be much smaller — depending on the agent's predictions regarding future events, adversary's prior belief about her, the types and strategies of other agents, and the random choices of the mechanism and of other agents. As discussed above, these can be hard to model, so it is tempting to use a worst-case model like differential privacy.

But what happens if we can formulate the agent's future transactions? What if we know that the agent is concerned with the belief of a specific adversary, and we can quantify the effects of changes to that belief? Is the behavior of a classical selfish agent in that case well-modeled by such a "DP-hardwired" privacy-aware agent? Will she even randomize her strategy? In other words, we ask:

What is the behavior of a selfish utility-maximizing agent in a setting with clear privacy costs?

More specifically, we ask whether we can take the above-mentioned toy-game and alter it by introducing payments between A and B such that the behavior of a privacy-aware agent in the toy-game matches the behavior of classical (non-privacy aware) agent in the altered game. In particular, in case B takes a randomized strategy — does her behavior preserve ϵ-differential privacy, and for what value of ϵ? The study of these questions may also provide insights relevant for traditional, non-game-theoretic uses of differential privacy — helping us understand how tightly differential privacy addresses the concerns of data subjects, and thus providing guidance in the setting of the privacy parameter ϵ or the use of alternative, non-worst-case variants of differential privacy (such as [1]).

Our model. In this work we consider multiple games that model an interaction between an agent which has a secret type and an adversary whose goal is to discover this type. Though the games vary in the resulting behavior of the agents, they all follow a common outline which is similar to the toy game mentioned above. Agent A offers B a free coupon, that comes in one of two types $\{0, 1\}$. Agent B has a secret type $t \in \{0, 1\}$ chosen from a known prior (D_0, D_1), such that a type-t agent has positive utility ρ_t for type-t coupon and zero utility for a type-$(1 - t)$ coupon. And so the game starts with B sending A a signal $\hat{t}$ indicating the requested type of coupon. (Formally, B's utility for the coupon is $\rho_t \mathbb{1}_{[\hat{t}=t]}$ for some parameters ρ_0, ρ_1.) Following this interaction, agent C, who viewed the signal $\hat{t}$ that B sent, challenges B into a game — with C taking action $\tilde{t}$ and incurring a payment from B of $P(\tilde{t}, t)$. To avoid the need to introduce a third party into the game, we identify C with A.[1] Figure 1 gives a schematic representation of the game's outline.

We make a few observations of the above interaction. We aim to model a scenario where B has the most incentive to hide her true type whereas A has the most incentive to discover B's type. Therefore, all of the payments we consider have the property that if B's type is t^* then $t^* = \arg\max_{\tilde{t}} P(\tilde{t}, t^*)$. Furthermore, the game is modeled so that the payments are transferred from B to A, which makes A's and B's goals as opposite as possible. (In fact, past the stage where B sends a signal $\hat{t}$, we have that A and B plays a zero-sum game.) We also note that A and B play a Bayesian game (in extensive form) as A doesn't know the private type of B, only its prior distribution. We characterize Bayesian Nash Equilibria (BNE) in this paper and will show that in each game, the BNE is unique except when parameters of the game satisfy certain equality constraints. It is not difficult to show that the strategies at every BNE of our games are part of a Perfect Bayesian Equilibrium (PBE), i.e. a subgame-perfect refinement of the BNE. However, we focus on BNE in this paper as the equilibrium refinement doesn't bring any additional insight to our problem.

[1] Hence the reason for the name "The Coupon Game". We think of A as G – an "evil" car-insurance company that offers its client a coupon either for an eyewear store or for a car race; thereby increasing the client's insurance premium based on either the client's bad eyesight or the client's fondness for speedy and reckless driving.

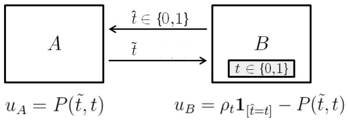

Fig. 1. A schematic view of the privacy game we model.

Our results and paper organization. First, in Section 2, following preliminaries we discuss the DP-hardwired privacy-aware agent as defined by Ghosh and Roth [13] and analyze her behavior in our toy game. Our analysis shows that given sufficiently large coupon valuations ρ_t, both types of B agent indeed play Randomized Response. We also discuss conditions under which other models of DP-hardwired privacy-aware agents play a randomized strategy.

Following preliminaries, we consider three different games. These games follow the general coupon-game outline, yet they vary in their payment function. The discussion for each of the games follows a similar outline. We introduce the game, then analyze the two agents' BNE strategies and see if the strategy of the B agent is indeed randomized or pure (and in case it is randomized — whether or not it follows Randomized Response for some value of ϵ). We also compare the coupon game to a "benchmark game" where B takes no action and A guesses B's type without any signal from B. Investigating whether it is even worth while for A to offer such a coupon, we compare A's profit between the two games.[2] The payment functions we consider are the following.

1. In Section 3 we consider the case where the payment function is given by a *proper scoring rule*. Proper scoring rules allows us to quantify the B's cost to any change in A's belief about her type. We show that in the case of symmetric scoring rules (scoring rules that are invariant to relabeling of event outcomes) both types of B agent follow a randomized strategy that causes A's posterior belief on the types to resemble Randomized Response. That is, initially A's belief on B being of type-0 (resp. type-1) is D_0 (resp. D_1); but B plays in a way such that after viewing the $\hat{t} = 0$ signal, A's belief that B is of type-0 (resp. type-1) is $\frac{1+\epsilon}{2}$ (resp. $\frac{1-\epsilon}{2}$) for some value of ϵ (and vice-versa in the case of the $\hat{t} = 1$ signal with the same ϵ).
2. In Section 4 we consider the case where the payments between A and B are the result of A guessing correctly B's type. A views the signal $\hat{t}$ and then guesses a type $\tilde{t} \in \{0, 1\}$ and receives a payment of $\mathbb{1}_{[\tilde{t}=t]}$ from B. This payment models the following viewpoint of B's future losses: there is a constant gap (of one "unit of utility") between interacting with an agent that knows B's type to an agent that does not know her type. We show that in this case, if the coupon valuations are fixed as ρ_0 and ρ_1, then at least one type of B agent plays deterministically. However, if B's valuation for the coupon is

[2] The benchmark game is not to be confused with the toy-game we discussed earlier in this introduction. In the toy game, A takes no action and B decides on a signal. In the benchmark game, B takes no action and A decides which action to take based on the specific payment function we consider in each game.

sampled from a continuous distribution, then A's strategy effectively dictates a threshold with the following property: any B agent whose valuation for the coupon is below the threshold lies and signals $\hat{t} = 1 - t$, and any agent whose valuation is above the threshold signals truthfully $\hat{t} = t$. Hence, an A agent who does not know B's valuation thinks of B as following a randomized strategy.

3. In Section 5 we consider a variation of the previous game where A also has the option to opt out and not challenge B into a payment game — to report $\perp$ and in return get no payment (i.e., $P(\perp, t) = 0$). We show that in such a game, under a very specific setting of parameters, the only BNE is such where both types of B agent take a randomized strategy. Under alternative settings of the game's parameters, the strategy of B is such that at least one of the two types plays deterministically.

Future directions are deferred to the full version of the paper, due to space limitation. We find it surprising to see how minor changes to the privacy payments lead to diametrically different behaviors. In particular, we see the existence of a threshold phenomena. Under certain parameter settings in the game we consider in item 3 above, we have that if the value of the coupon is above a certain threshold then at least one of the two types of B agent plays deterministically; and if the value of the coupon is below this threshold, B randomizes her behavior s.t. $\hat{t} = t$ w.p. close to $\frac{1}{2}$.

1.1 Related Work

The study of the intersection between mechanism design and differential privacy began with the seminal work of McSherry and Talwar [18], who showed that an ϵ-differentially private mechanism is also ϵ-truthful. The first attempt at defining a privacy-aware agent was of Ghosh and Roth [13] who quantified the privacy loss using a linear approximation $v_i \cdot \epsilon$ where v_i is an individual parameter and ϵ is the level of differential privacy that a mechanism preserves. Other applications of differentially privacy mechanisms in game theoretic settings were studied by Nissim et al [20]. The work of Xiao [24] initiated the study of mechanisms that are truthful even when you incorporate the privacy loss into the agents' utility functions. Xiao's original privacy loss measure was the mutual information between the mechanism's output and the agent's type. Nissim et al [19] (who effectively proposed a preliminary version of our coupon game called "Rye or Wholewheat") generalized the models of privacy loss to only assume that it is *upper bounded* by $v_i \cdot \epsilon$. Chen et al [4] proposed a refinement where the privacy loss is measured with respect to the given input and output. Fleischer and Lyu [11] considered the original model of agents as in Ghosh and Roth [13] but under the assumption that v_i, the value of the privacy parameter of each agent, is sampled from a known distribution.

Several papers in economics look at the potential loss of agents from having their personal data revealed. In fact, one folklore objection to the Vickrey auction is that in a repeated setting, by providing the sellers with the bidders' true valuations for the item, the bidders subject themselves to future loss should the seller prefer to run a reserved-price mechanism in the future. In the context of repeated interaction between an agent and a company, there have been works [6,2] studying the effect of price differentiation based on an agent allowing the company to remember whether she purchased the same item in the past. Interestingly, strategic agents realize this effect and so they might "haggle" — reject a price below their valuation for the item in round 1 so that they'd be able to get even lower price in round 2. In that sense, the fact that the agents publish their past interaction with the company actually helps the agents. Other work [3] discusses a setting where a buyer sequentially interacts with two different sellers, and characterizes the conditions under which the first seller prefers not to give the buyer's information to the second seller. Concurrently with our work, Gradwohl and Smorodinsky [15], whose motivation is to analyze the effect of privacy concerns, introduce a framework of games in which an agent's utility is affected by both her actions and how her actions are perceived by a third party.

The privacy games that we propose and analyze in this paper fall into the class of signaling games [17], where a sender (B in our game) with a private type sends a message (i.e. a signal) to a receiver (A in our game) who then takes an action. The payoffs of both players depend on the sender's message, the receiver's action, and the sender's type. Signaling games have been widely used in modeling behavior in economics and biology. The focus is typically on understanding when signaling is informative, i.e. when the message of the sender allows the receiver to infer the sender's private type with certainty, especially in settings when signaling is costly (e.g. Spence's job market signaling game [21]). In our setting, however, informative signaling violates privacy. We are interested in characterizing when the sender plays in a way such that the receiver cannot infer her type deterministically.

2 Preliminaries

2.1 Equilibrium Concept

We model the games between A and B as Bayesian extensive-form games. However, instead of using the standard Perfect Bayesian Equilibrium (PBE), which is a refinement of Bayesian Nash Equilibrium (BNE) for extensive-form games, as our solution concept, we analyze BNE for our games. It can be shown that all of the BNEs considered in our paper can be "extended" to PBEs (by appropriately defining the beliefs of agent A about agent B at all points in the game). We thus avoid defining the more subtle concept of PBE as the refinement doesn't provide additional insights for our problem. Below we define BNE.

A *Bayesian* game between two agents A and B is specified by their type spaces $(\mathit{\Gamma}_A, \mathit{\Gamma}_B)$, a prior distribution $\mathit{\Pi}$ over the type spaces (according to which nature draws the private types of the agents), sets of available actions (C_A, C_B), and utility functions, $u_i : \mathit{\Gamma}_A \times \mathit{\Gamma}_B \times C_A \times C_B \to \mathbb{R}$, $i \in \{A, B\}$. A *mixed* or *randomized* strategy of agent i maps a type of agent i to a distribution over her available actions, i.e. $\sigma_i : \mathit{\Gamma}_i \to \mathit{\Delta}(C_i)$, where $\mathit{\Delta}(C_i)$ is the probability simplex over C_i.

Definition 1. *A strategy profile* (σ_A, σ_B) *is a* Bayesian Nash Equilibrium *if*

$$\mathbf{E}[u_i(T_i, T_{-i}, \sigma_i(T_i), \sigma_{-i}(T_{-i}))|T_i = t_i] \geq \mathbf{E}[u_i(T_i, T_{-i}, \sigma'_i(T_i), \sigma_{-i}(T_{-i}))|T_i = t_i]$$

for all $i \in \{A, B\}$*, all types* $t_i \in \Gamma_i$ *occurring with positive probability, and all strategies* σ'_i*, where* σ_{-i} *and* T_{-i} *denote the strategy and type of the other agent respectively and the expectation is taken over the randomness of agent type* T_{-i} *and the randomness of the strategies,* σ_i*,* σ_{-i} *and* σ'_i*.*

2.2 Differential Privacy

In order to define differential privacy, we first need to define the notion of neighboring inputs. Inputs are elements in $\mathcal{X}^n$ for some set $\mathcal{X}$, and two inputs $\mathcal{I}, \mathcal{I}' \in \mathcal{X}^n$ are called neighbors if the two are identical on the details of all individuals (all coordinates) except for at most one.

Definition 2 ([9]). *An algorithm* ALG *which maps inputs into some range* $\mathcal{R}$ *satisfies* ϵ-differential privacy *if for all pairs of neighboring inputs* $\mathcal{I}, \mathcal{I}'$ *and for all subsets* $\mathcal{S} \subset \mathcal{R}$ *it holds that* $\mathbf{Pr}[\mathsf{ALG}(\mathcal{I}) \in \mathcal{S}] \leq e^{\epsilon}\mathbf{Pr}[\mathsf{ALG}(\mathcal{I}') \in \mathcal{S}]$.

One of the simplest algorithms that achieve ϵ-differential privacy is called *Randomized Response* [16,10], which dates back to the 60s [22]. This algorithm is best illustrated over a binary input, where each individual is represented by a single binary bit (therefore a neighboring instance is a neighbor in which one individual is represented by a different bit), Randomized Response works by perturbing the input. For each individual i represented by the bit b_i, the algorithm randomly and independently picks a bit $\hat{b}_i$ s.t. $\mathbf{Pr}[\hat{b}_i = b_i] = \frac{1+\epsilon}{2}$ for some $\epsilon \in [0, 1)$. It follows from the definition of the algorithm that it satisfies $\ln(\frac{1+\epsilon}{1-\epsilon}) \approx 2\epsilon$-differential privacy. Randomized Response is sometimes presented as a distributed algorithm, where each individual randomly picks $\hat{b}_i$ locally, and reports $\hat{b}_i$ publicly. Therefore, it is possible to view this work as an investigation of the type of games in which selfish utility-maximizing agents truthfully follow Randomized Response, rather than sending some arbitrary bit as $\hat{b}_i$.

In this work, we define certain games and analyze the behavior of the two types of B agent in the BNE of these games. And so, denoting B's strategy as σ_B, we consider the implicit algorithm $\sigma_B(t)$ that tells a type-t agent what probability mass to put on the 0-signal and on the 1-signal. Knowing B's strategy σ_B, we say that B satisfies $\ln(X_{\text{game}})$-differential privacy where[3]

$$X_{\text{game}} \overset{\text{def}}{=} X_{\text{game}}(\sigma_B) = \max_{t, \hat{t} \in \{0,1\}} \left(\frac{\mathbf{Pr}[\sigma_B(t) = \hat{t}]}{\mathbf{Pr}[\sigma_B(1-t) = \hat{t}]} \right)$$

We are interested in finding settings where $X_{\text{game}}(\sigma^*_B)$ is finite, where σ^*_B denotes B's BNE strategy. We say B *plays a Randomized Response strategy* in a game whenever her BNE strategy σ^*_B satisfies $\mathbf{Pr}[\sigma^*_B(0) = 0] = \mathbf{Pr}[\sigma^*_B(1) = 1] = p$ for some $p \in [1/2, 1)$.

[3] We use the convention $\frac{0}{0} = 1$.

Privacy-Aware Agents. The notion of privacy-aware agents has been developed through a series of works [24,13,19,4]. The utility function of our privacy-aware agent B is of the form $u_B = u_B^{out} - u_B^{priv}$. The first term, u_B^{out} is the utility of agent B from the mechanism. The second term, u_B^{priv}, represents the agent's privacy loss. The exact definition of u_B^{priv} (and even the variables u_B^{priv} depends on) varies between the different works mentioned above, but all works bound the privacy-loss of an agent that interacts with a mechanism that satisfies ϵ-differential privacy by $u_B^{priv} \leq v \cdot \epsilon$ for some $v > 0$. Here we argue about the behavior of a privacy-aware agent with the maximal privacy loss function, which is the type of agent considered by Ghosh and Roth [13] (i.e., the agent's privacy loss when interacting with a mechanism that satisfies ϵ-differential privacy is exactly $v \cdot \epsilon$ for some $v > 0$).

Recall our toy game: B sends a signal $\hat{t}$ and gets a coupon of type $\hat{t}$. Therefore, the outcome of this simple game is $\hat{t}$, precisely the action that B takes. B's type is picked randomly to be 0 w.p. D_0 and 1 w.p. D_1, and a B agent of type t has valuation of ρ_t for a coupon of type t. Therefore, in this game $u_B^{out} = \rho_t \mathbb{1}_{[\hat{t}=t]}$. The mechanism we consider is σ_B^*, B's utlity-maximizing strategy, which we think of as the implicit algorithm that tells a type-t agents what probability mass to put on sending the $\hat{t} = 0$ signal and what mass to put on the $\hat{t} = 1$ signal. As noted above, this strategy satisfies $\ln(X_{\text{game}})$-differential privacy, and so $u_B^{priv}(\sigma_B^*) = v \cdot \ln(X_{\text{game}})$ for some parameter $v > 0$. Assuming $D_0\rho_0 \neq D_1\rho_1$, our proof shows that this privacy-aware agent chooses essentially between two alternatives in our toy game: either both types take the same deterministic strategy and send the same signal ($\mathbf{Pr}[\sigma_B^*(0) = b] = \mathbf{Pr}[\sigma_B^*(1) = b] = 1$ for some $b \in \{0, 1\}$); or the agent randomizes her behavior and plays using Randomized Response: $\mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1] \in [\frac{1}{2}, 1)$. We show that for sufficiently large values of the coupon the latter alternative is better than the first.

Theorem 3. *Let B be a privacy-aware agent, whose privacy loss is given by $v \ln(X_{\text{game}})$ for some $v > 0$. Assume that there exists an $\alpha > 0$ s.t. for sufficiently large values of ρ_0, ρ_1 it holds that $\min\{\rho_0, \rho_1\} \geq \alpha \cdot (\rho_0 + \rho_1)$. Then, the unique strategy σ_B^* that maximizes B's utility is randomized and satisfies: $\mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1] = p^*$ for some $p^* \in [\frac{1}{2}, 1)$.*

The proof is deferred to the full version of the paper. The proof of Theorem 3 also applies to some alternative models of a privacy-aware agent. In addition to Theorem 3, we also analyze, for completeness, an alternative scenario where type 0 and type 1 are two competing agents. Observe that this is no longer a Bayesian game with a single player but rather a standard complete-information game with two players. We show that this game also has NEs where both types play randomized strategies that follow Randomized Response (i.e.,$\mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1]) > \frac{1}{2}$).

3 The Coupon Game with Scoring Rules Payments

In this section, we model the payments between A and B using a proper scoring rule (see below). This model is a good "first-attempt" model for the following two reasons. (i) Proper scoring rules assign profit to A based on the accuracy of her belief, so A has incentives to improve her prior belief on B's type. (ii) As we show, in this model it is possible to quantify the B's trade-off between an ϵ-change in the belief and the cost that B pays A. In that aspect, this model gives a clear quantifiable trade-off that explains what each additional unit of ϵ-differential privacy buys B. Interestingly, proper scoring rules were recently applied in the context of differential privacy [12] (yet in a very different capacity).

Proper scoring rules (see surveys [23,14]) were devised as a method to elicit experts to report their true prediction about some random variable. For a $\{0,1\}$-valued random variable X, an expert is asked to report a prediction $x \in [0,1]$ about the probability that $X = 1$. We pay her $f_1(x)$ if indeed $X = 1$ and $f_0(x)$ otherwise. A *proper scoring rule* is a pair of functions (f_0, f_1) such that $\arg\max_x \mathbf{E}_{t \leftarrow X}[f_t(x)] = \mathbf{Pr}[X = 1]$. Hence a risk-neutral agent's best strategy is to report $x = \mathbf{Pr}[X = 1]$. Most frequently used proper scoring rules are *symmetric* (or label-invariant) rules, where $\forall x, f_1(x) = f_0(1 - x)$ (also referred to as neutral scoring rules in [5]). With symmetric proper scoring rules, the payment to an expert reporting x as the probability of a random variable X to be 1, is identical to the payment of an expert reporting $(1-x)$ as the probability of the random variable $(1-X)$ to be 1. Additional background regarding proper scoring rules is deferred to the full version of this paper.

3.1 The Game with Scoring Rule Payments

We now describe the game, and analyze its BNE. In this game A interacts with a random B from a population that has D_0 fraction of type 0 agents and D_1 fraction of type 1 agents. Wlog we assume throughout Sections 3, 4 and 5 that $D_0 \geq D_1$. A aims to discover B's secret type. She has utility that is directly linked to her posterior belief on B's type and A reports her belief that B is of type 1. A's payments are given by a proper scoring rule, composed of two functions (f_0, f_1), so that after reporting a belief of x, a B agent of type t pays $f_t(x)$ to A.

A benchmark game. First consider the following straight forward (and more boring) game where B does nothing, A merely reports x – her belief that B is of type 1. In this game A gets paid according to a proper scoring rule — i.e., A gets a payment of $F_{D_1}(x) \overset{\text{def}}{=} D_0 f_0(x) + D_1 f_1(x)$ in expectation. Since (f_0, f_1) is a proper scoring rule, A maximizes her expected payment by reporting $x = D_1$. So, in this game A gets paid $g(D_1) \overset{\text{def}}{=} f_{D_1}(D_1)$ in expectation, whereas B's expected cost is $g(D_1)$. (Alternatively, a B agent of type 0 pays $f_0(D_1)$ and a B agent of type 1 pays $f_1(D_1)$.)

The full game. We now turn our attention to a more involved game. Here A, aiming to have a more accurate posterior belief on B's type, offers B a coupon. Agents of type t prefer a coupon of type t. And so, B chooses what type to report A, who then gives B the coupon and afterwards makes a prediction about B's probability of being of type 1. The formal stages of the game are as follows.

0. B's type, t, is drawn randomly with $\mathbf{Pr}[t = 0] = D_0$ and $\mathbf{Pr}[t = 1] = D_1$.
1. B reports to A a type $\hat{t} = \sigma_B(t)$ and receives utility of ρ_t if indeed $\hat{t} = t$. We assume throughout this section that $\rho_0 = \rho_1 = \rho$.

2. A reports a prediction x, representing $\mathbf{Pr}[t = 1 \mid \sigma_B(t) = \hat{t}]$, and receives a payment from B of $f_t(x)$.

Theorem 4. *Consider the coupon game with payments in the form of a symmetric proper scoring rule and with the following added assumption about the value of the coupon: $f_1(D_0) - f_1(D_1) < \rho < f_1(1) - f_1(0) = f_0(0) - f_0(1)$. The unique BNE strategy of B in this game, denoted σ^*_B, satisfies that $\mathbf{Pr}[t = 0 \mid \sigma^*_B(t) = 0] = \mathbf{Pr}[t = 1 \mid \sigma^*_B(t) = 1]$.*

Note that a Randomized Response strategy σ_B for B would instead have $\mathbf{Pr}[\sigma_B(0) = 0] = \mathbf{Pr}[\sigma_B(1) = 1]$. This condition is different from the condition in Theorem 4 when $\mathbf{Pr}[t = 0] \neq \mathbf{Pr}[t = 1]$ (i.e., $D_0 \neq D_1$). The proof of Theorem 4 is in the full version of this paper, where we also compare A's profit in the benchmark game to her profit from her BNE strategy in the full game.

4 The Coupon Game with the Identity Payments

In this section, we examine a different variation of our initial game. As always, we assume that B has a type sampled randomly from $\{0, 1\}$ w.p. D_0 and D_1 respectively, and wlog $D_0 \geq D_1$. Yet this time, the payments between A and B are given in the form of a 2×2 matrix we denote as M. This payment matrix specifies the payment from B to A in case A "accuses" B of being of type $\tilde{t} \in \{0, 1\}$ and B is of type t. In general we assume that A strictly gains from finding out B's true type and potentially loses otherwise (or conversely, that a B agent of type t strictly loses utility if A accuses B of being of type $\tilde{t} = t$ and potentially gains money if A accuses B of being of type $\tilde{t} = 1 - t$). In this section specifically, we consider one simple matrix M – the identity matrix $I_{2\times2}$. Thus, A gets utility of 1 from correctly guessing B's type (the same utility regardless of B's type being 0 or 1) and 0 utility if she errs.

4.1 The Game and Its Analysis

The benchmark game. The benchmark for this work is therefore a very simple "game" where B does nothing, A guesses a type and B pays A according to M. It is clear that A maximizes utility by guessing $\tilde{t} = 0$ (since $D_0 \geq D_1$) and so A gains in expectation D_0; where an agent B of type $t = 0$ pays 1 to A, and an agent B of type $t = 1$ pays 0 to A.

The full game. Aiming to get a better guess for the actual type of B, we now assume A first offers B a coupon. As before, B gets a utility of ρ_t from a coupon of the right type and 0 utility from a coupon of the wrong type. And so, the game takes the following form now.

0. B's type, denoted t, is chosen randomly, with $\mathbf{Pr}[t = 0] = D_0$ and $\mathbf{Pr}[t = 1] = D_1$.
1. B reports a type $\hat{t} = \sigma_B(t)$ to A. A in return gives B a coupon of type $\hat{t}$.
2. A accuses B of being of type $\tilde{t} = \sigma_A(\hat{t})$ and B pays 1 to A if indeed $\tilde{t} = t$.

And so, the utility of agent A is $u_A = \mathbb{1}_{[\tilde{t}=t]}$. The utility of agent B is a summation of two factors – reporting the true type to get the right coupon and the loss of paying A for finding B's true type. So $u_B = \rho_t \mathbb{1}_{[\hat{t}=t]} - \mathbb{1}_{[\tilde{t}=t]}$.

Theorem 5. *In the coupon game with payments given by the identity matrix with $\rho_0 \neq \rho_1$, any BNE strategy of B is pure for at least one of the two types of B agent. Formally, for any BNE strategy of B, denoted σ_B^*, there exist $t, \hat{t} \in \{0,1\}$ s.t. $\mathbf{Pr}[\sigma_B^*(t) = \hat{t}] = 1$.*

In the case where $\rho_0 = \rho_1$ then B has infinitely many randomized BNE strategies, including a BNE strategy σ_B^* s.t. $\frac{1}{2} \leq \mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1] < 1$ (Randomized response).

4.2 Continuous Coupon Valuations

We now consider the same game with the same payments, but under a different setting. Whereas before we assumed the valuations that the two types of B agents have for the coupon are fixed (and known in advance), we now assume they are not fixed. In this section we assume the existence of a continuous prior over ρ, where each type $t \in \{0,1\}$ has its own prior, so $\mathsf{CDF}_0(x) \overset{\text{def}}{=} \mathbf{Pr}[\rho < x \mid t = 0]$ with an analogous definition of $\mathsf{CDF}_1(x)$. We use CDF_B to denote the cumulative distribution function of the prior over ρ (i.e., $\mathsf{CDF}_B(x) = \mathbf{Pr}[\rho < x] = D_0\mathsf{CDF}_0(x) + D_1\mathsf{CDF}_1(x)$). We assume the CDF is continuous and so $\mathbf{Pr}[\rho = y] = 0$ for any y. Given any $z \geq 0$ we denote $\mathsf{CDF}_B^{-1}(z)$ the set $\{y : \ \mathsf{CDF}_B(y) = z\}$.

Theorem 6. *In every BNE (σ_A^*, σ_B^*) of the coupon game with identity payments, where $D_0 \neq D_1$ and the valuations of the B agents for the coupon are taken from a continuous distribution over $[0, \infty)$, the BNE-strategies are as follows.*

- *Agent A always plays $\tilde{t} = 0$ after viewing the $\hat{t} = 0$ signal (i.e., $\mathbf{Pr}[\sigma_A^*(0) = 0] = 1$); and plays $\tilde{t} = 1$ after viewing the $\hat{t} = 1$ signal with probability y^* (i.e., $\mathbf{Pr}[\sigma_A^*(1) = 1] = y^*$), where y^* is any value in $\mathsf{CDF}_B^{-1}(D_1)$ when $\mathbf{Pr}[\rho < 1] \geq D_1$ and $y^* = 1$ when $\mathbf{Pr}[\rho < 1] < D_1$.*
- *Agent B reports truthfully (sends the signal $\hat{t} = t$) whenever her valuation for the coupon is greater than y^*, and lies (sends the signal $\hat{t} = 1 - t$) otherwise. That is, for every $t \in \{0,1\}$ and $\rho \in [0, \infty)$, we have that if $\rho > y^*$ then $\mathbf{Pr}[\sigma_B^*(t) = t] = 1$ and if $\rho < y^*$ then $\mathbf{Pr}[\sigma_B^*(t) = t] = 0$.*

Due to space constraints, this analysis is deferred to the full version of the paper.

5 The Coupon Game with an Opt Out Strategy

In this section, we consider a version of the game considered in Section 4. The revised version of the game we consider here is very similar to the original game, except for A's ability to "opt out" and not guess B's type.

In this section, we consider the most general form of matrix payments. We replace the identity-matrix payments with general payment matrix M of the form $M = \left[\begin{array}{c|c} M_{0,0} & -M_{0,1} \\ \hline -M_{1,0} & M_{1,1} \end{array}\right]$ with the (i,j) entry in M means A guessed $\tilde{t} = i$ and B's true type is $t = j$, and so B pays A the amount detailed in the (i,j)-entry. We assume $M_{0,0}, M_{0,1}, M_{1,0}, M_{1,1}$ are all non-negative.

Indeed, when previously considering the identity matrix payments, we assumed the for A, realizing that B has type $t = 0$ is worth just as much as finding B has type $t = 1$. But it might be the case that finding a person of $t = 1$ should be more worthwhile for A. For example, type $t = 1$ (the minority, since we always assume $D_0 \geq D_1$) may represent having some embarrassing medical condition while type $t = 0$ representing not having it. Therefore, $M_{1,1}$ can be much larger than $M_{0,0}$, but similarly $M_{1,0}$ is probably larger than $M_{0,1}$. (Falsely accusing B of being of the embarrassing type is costlier than falsely accusing a B of type 1 of belonging to the non-embarrassing majority.) Our new payment matrix still motivates A to find out B's true type — A gains utility by correctly guessing B's type, and loses utility by accusing B of being of the wrong type.

The "strawman" game. First, consider a simple game where B makes no move (A offers no coupon) and A tries to guess B's type without getting any signal from B. Then A has three possible pure strategies: (i) guess that B is of type 0; (ii) guess that B is of type 1; and (iii) guess nothing. In expectation, the outcome of option (i) is $D_0 M_{0,0} - D_1 M_{0,1}$ and the outcome of option (ii) is $D_1 M_{1,1} - D_0 M_{1,0}$. If the parameters of M are set such that both options are negative then A's preferred strategy is to opt out and gain 0. We assume throughout this section that indeed the above holds. (Intuitively, this assumption reflects the fact that we don't make assumptions about people's type without first getting any information about them.) So we have

$$\frac{M_{0,0}}{M_{0,1}} < \frac{D_1}{D_0} , \qquad \text{and} \qquad \frac{M_{1,1}}{M_{1,0}} < \frac{D_0}{D_1} \tag{1}$$

A direct (and repeatedly used) corollary of Equation (1) is that $\frac{M_{0,0}}{M_{0,1}} < \frac{M_{1,0}}{M_{1,1}}$.

The full game. We now give the formal description of the game.

0. B's type, denoted t, is chosen randomly, with $\mathbf{Pr}[t = 0] = D_0$ and $\mathbf{Pr}[t = 1] = D_1$.
1. B reports a type $\hat{t}$ to A. A in return gives B a coupon of type $\hat{t}$.
2. A chooses whether to accuse B of being of a certain type, or opting out.
   - If A opts out (denoted as $\tilde{t} = \perp$), then B pays A nothing.
   - If A accuses B of being of type $\tilde{t}$ then: if $\tilde{t} = t$ then B pays $M_{t,t}$ to A, and if $\tilde{t} = 1 - t$ then B pays $-M_{1-t,t}$ to A (or A pays $M_{1-t,t}$ to B).

Introducing the option to opt out indeed changes significantly the BNE strategies of A and B.

Theorem 7. *If we have that $D_0^2 M_{0,0} M_{1,0} = D_1^2 M_{0,1} M_{1,1}$ and the parameters of the game satisfy the following condition:*

$$\begin{aligned} 0 < \rho_1 M_{1,0} - \rho_0 M_{1,1} < M_{0,1} M_{1,0} - M_{0,0} M_{1,1} \\ 0 < \rho_0 M_{0,1} - \rho_1 M_{0,0} < M_{0,1} M_{1,0} - M_{0,0} M_{1,1} \end{aligned} \tag{2}$$

then the unique BNE strategy of B, denote σ_B^*, is such that B plays Randomized Response: $\frac{1}{2} \leq \mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1] < 1$.

Proving Theorem 7 is the goal of this section. The proof itself is deferred to the full version of this paper, where we also give a complete summary of the various BNEs of this game. We detail 6 different cases that cover all possible settings of the game. Each of these 6 cases is defined by a different *feasibility* condition. These conditions guarantee that A is able to find a strategy that cause at least one of the two types of B agent to be indifferent as to the signal she sends.

The feasibility condition detailed in Equation (2) can be realized starting with any matrix M satisfying $M_{0,0}M_{1,1} < M_{0,1}M_{1,0}$ (which is a necessary condition derived from Equation (1)), which intuitively can be interpreted as having a wrong "accusation" being costlier than the gain from a correct "accusation" (on average and in absolute terms). Given such M, one can set D_0 and D_1 s.t. $\frac{D_0}{D_1} = \sqrt{\frac{M_{0,1}}{M_{0,0}} \cdot \frac{M_{1,1}}{M_{1,0}}}$ as to satisfy Equation (1). This can be interpreted as balancing the "significance" of type 0 (i.e. $M_{0,0}M_{0,1}$) with the "significance" of type 1 (i.e. $M_{1,0}M_{1,1}$), setting the more significant type as the less probable (i.e. if type 1 is more significant than type 0, than $D_1 < D_0$). We then pick ρ_0, ρ_1 that satisfy $\frac{M_{1,1}}{M_{1,0}} < \frac{\rho_1}{\rho_0} < \frac{M_{0,1}}{M_{0,0}}$ and scale both by the sufficiently small multiplicative factor so we satisfy the other inequality in Equation (2). (In particular, setting $\frac{\rho_1}{\rho_0} = \frac{D_0}{D_1}$ is a feasible solution.) Here, ρ_0 and ρ_1 are set such that the ratio $\frac{\rho_1}{\rho_0}$ balances the significance ratio w.r.t type 1 accusation (i.e. $\frac{\rho_1}{\rho_0} > \frac{M_{1,1}}{M_{0,0}}$) and the ratio $\frac{\rho_0}{\rho_1}$ balances the significance ratio w.r.t to type 0 accusation (i.e. $\frac{\rho_0}{\rho_1} > \frac{M_{0,0}}{M_{1,0}}$). More concretely, for any matrix $M = \begin{pmatrix} 1 & c \\ c & d \end{pmatrix}$ with parameters c, d satisfying $d < c^2$, we can set $\frac{D_0}{D_1} = \sqrt{d}$ and any sufficiently small ρ_0, ρ_1 satisfying $\frac{\rho_1}{\rho_0} \in (\frac{d}{c}, c)$ and satisfy the requirements of Theorem 7.

Recall, in addition to the conditions specifically stated in Equation (2), we also require that $D_0^2 M_{0,0}M_{1,0} = D_1^2 M_{0,1}M_{1,1}$ in order for the two types of agent B to play Randomized Response. In other words, the feasibility condition in Equation (2) implies that B's BNE strategy, denoted by $p^* = \mathbf{Pr}[\sigma_B^*(0) = 0]$ and $q^* = \mathbf{Pr}[\sigma_B^*(1) = 1]$, is given by

$$(p^*, q^*) = (\frac{D_0D_1M_{0,1}M_{1,0} - D_1^2 M_{0,1}M_{1,1}}{D_0D_1M_{0,1}M_{1,0} - D_0D_1M_{0,0}M_{1,1}}, \frac{D_0D_1M_{0,1}M_{1,0} - D_0^2 M_{0,0}M_{1,0}}{D_0D_1M_{0,1}M_{1,0} - D_0D_1M_{0,0}M_{1,1}})$$

The additional condition of $D_0^2 M_{0,0}M_{1,0} = D_1^2 M_{0,1}M_{1,1}$ implies therefore that $p^* = q^*$. And so, in this case the B agent plays a Randomized Response strategy that preserves ϵ-differential privacy for $\epsilon = \ln(\frac{p^*}{1-q^*}) = \ln\left(\frac{D_1 M_{0,1}}{D_0 M_{0,0}}\right)$. Observe that this value of ϵ is *independent* from the value of the coupon (i.e., from ρ_0 and ρ_1). This is due to the nature of BNE in which an agent plays her Nash-strategy in order to make her opponent indifferent between various strategies rather than maximizing her own utility. Therefore, the coordinates (p^*, q^*) are such that they make agent A indifferent between several pure strategies. And since the utility function of A is independent of ρ_0, ρ_1, we have that perturbing the values of ρ_0, ρ_1 does not affect the coordinates (p^*, q^*). (Yet, perturbing the values of ρ_0, ρ_1 does affect the various relations between the parameters of the game, and so it may determine which of the 6 feasibility conditions does in fact hold.)

Acknowledgments

We would like to thank Kobbi Nissim for many helpful discussions and helping us in initiating this line of work.

References

1. Bassily, R., Groce, A., Katz, J., Smith, A.: Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In: FOCS (2013)
2. Bergemann, D., Brooks, B., Morris, S.: The Limits of Price Discrimination. Cowles Foundation Discussion Papers 1896, Cowles Foundation for Research in Economics, Yale University (May 2013), `http://ideas.repec.org/p/cwl/cwldpp/1896.html`
3. Calzolari, G., Pavan, A.: On the optimality of privacy in sequential contracting. Journal of Economic Theory 130(1) (2006)
4. Chen, Y., Chong, S., Kash, I.A., Moran, T., Vadhan, S.P.: Truthful mechanisms for agents that value privacy. In: EC (2013)
5. Chen, Y., Devanur, N.R., Pennock, D.M., Vaughan, J.W.: Removing arbitrage from wagering mechanisms. In: EC (2014)
6. Conitzer, V., Taylor, C.R., Wagman, L.: Hide and seek: Costly consumer privacy in a market with repeat purchases. Marketing Science 31(2) (2012)
7. Dwork, C.: Differential privacy. In: ICALP (2006)
8. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: EUROCRYPT (2006)
9. Dwork, C., Mcsherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: TCC (2006)
10. Dwork, C., Smith, A.: Differential privacy for statistics: What we know and what we want to learn. Journal of Privacy and Confidentiality 1(2), 2 (2010)
11. Fleischer, L., Lyu, Y.H.: Approximately optimal auctions for selling privacy when costs are correlated with data. In: EC (2012)
12. Ghosh, A., Ligett, K., Roth, A., Schoenebeck, G.: Buying private data without verification. In: EC (2014)
13. Ghosh, A., Roth, A.: Selling privacy at auction. In: EC (2011)
14. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477), 359–378 (2007)
15. Gradwohl, R., Smorodinsky, R.: Subjective perception games and privacy. CoRR abs/1409.1487 (2014), `http://arxiv.org/abs/1409.1487`
16. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? In: FOCS (2008)
17. Mas-Colell, A., Whinston, M.D., Green, J.R.: Microeconomic Theory. Oxford University Press (Jun 1995)
18. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: FOCS. pp. 94–103 (2007)
19. Nissim, K., Orlandi, C., Smorodinsky, R.: Privacy-aware mechanism design. In: EC (2012)
20. Nissim, K., Smorodinsky, R., Tennenholtz, M.: Approximately optimal mechanism design via differential privacy. In: ITCS (2012)
21. Spence, M.: Job market signalling. Quarterly Journal of Economics 87(3), 355–374 (August 1973)
22. Warner, S.L.: Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association 60(309) (Mar 1965)
23. Winkler, R.: Scoring rules and the evaluation of probabilities. Test 5(1), 1–60 (1996)
24. Xiao, D.: Is privacy compatible with truthfulness? In: ITCS (2013)