Privacy Games

YILING CHEN, Harvard University, USA
OR SHEFFET*, Bar-Ilan University, Israel
SALIL VADHAN, Harvard University, USA

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 [4, 14, 20, 28]. 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.

CCS Concepts: • **Theory of computation** → *Theory and algorithms for application domains*; *Algorithmic game theory and mechanism design*; *Quality of equilibria*;

Additional Key Words and Phrases: Differential privacy, Bayes-Nash equilibrium, privacy modeling

ACM Reference format:
Yiling Chen, Or Sheffet, and Salil Vadhan. 2020. Privacy Games. *ACM Trans. Econ. Comput.* 8, 2, Article 9 (May 2020), 37 pages.
https://doi.org/10.1145/3381533

Previous versions of this article appear in the Proceedings of the 10th International Conference on Web and Internet Economics (WINE 2014), Beijing, China, and on arXiv:1410.1920 [cs.GT].
Y. Chen was Supported in part by NSF grant CCF-1301976.
*The bulk of the work was done when the author was a postdoctoral fellow at Harvard University, supported in part by NSF grant CNS-1237235.
S. Vadhan was Supported by NSF grant CNS-1237235, a gift from Google, Inc., and a Simons Investigator grant.
Authors' addresses: Y. Chen and S. Vadhan, School of Engineering and Applied Sciences, Harvard University, 33 Oxford St, Cambridge, MA 02138, USA; emails: yiling@eecs.harvard.edu, salil_vadhan@harvard.edu; O. Sheffet, Faculty of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel; email: or.sheffet@biu.ac.il.

2167-8375/2020/05-ART9 $15.00
https://doi.org/10.1145/3381533

1 INTRODUCTION

In recent years, as the subject of privacy has become 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.

Differential privacy (DP), a mathematical model for privacy, developed for statistical data analysis [8, 9], avoids the need for such intricate modeling by providing a worst-case bound on agents' exposure to privacy loss. Specifically, by using an 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 [14, 20], using an 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 [4, 14, 20, 28] 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 versus a (negative) loss from their nonprivate 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 [20].) Therefore, it is simple to see that a non-privacy-aware agent always (deterministically) asks for the gift that matches his or 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 [14] (and the value of the coupon is large enough), a privacy-aware agent takes a randomized strategy. (See Section 2.2.1.) Specifically, the agent plays *Randomized Response*, a standard differentially private mechanism that outputs a random choice slightly biased toward the agent's favorable action.

However, it was argued [4, 20] that it is not realistic to use the worst-case model of DP to quantify the agent's privacy loss and predict his or 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, the adversary's prior belief about him or 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

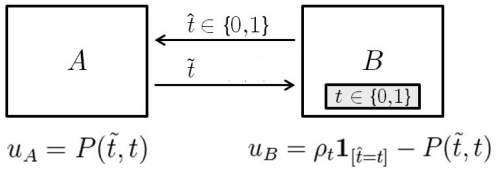

Fig. 1. A schematic view of the privacy game we model.

such a "DP-hardwired" privacy-aware agent? Will the agent even randomize his or 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 the classical (non-privacy-aware) agent in the altered game. In particular, in case B takes a randomized strategy, does his or 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 that 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 that is similar to the toy game mentioned above. Agent A offers B a free coupon, which 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 a 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.) Next, having observed the signal $\hat{t}$ that B sent, A challenges B—A takes an action $\tilde{t}$ and as a result B pays A a payment of $P(\tilde{t}, t)$.[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 is a consumer balancing his or her immediate needs (the reward from getting the coupon that best matches his or her type) with future privacy concerns (the payments to A). We thus picked a model in which 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 play a zero-sum game.) Note that the value of the coupon does not affect A's utility—since, again, our focus is on a model where A's incentive is to find out B's type. (Had we modeled A's utility as a function also of the coupon's value, then our focus would have shifted toward the study of whether discovering the type of B—or more broadly, whether such "privacy violation"—is worthwhile for A or not.) We also note that A and B play a Bayesian game (in extensive form) as A doesn't know the

[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.

private type of B, only its prior distribution. We characterize Bayesian Nash Equilibria (BNE) in this article 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, when "padded" with the appropriate posterior beliefs (see further discussion in Section 2).

Our results and article organization. First, in Section 2, following preliminaries, we discuss the DP-hardwired privacy-aware agent as defined by Ghosh and Roth [14] and analyze his or 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 worthwhile 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 allow us to quantify B's cost for any change in A's belief about his or her type; thus, if we view privacy concerns as a direct attempt to influence an adversary's belief regarding you, it is only natural we study a model where payment directly relates to an adversary's belief on the true type of B. 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—which is *not* Randomized Response, yet the purpose of B's BNE strategy is to cause A's posterior belief on the types to resemble Randomized Response. That is, initially A's belief on B being of type-0 (type-1, respectively) is D_0 (D_1, respectively), but B plays in a way such that after viewing the $\hat{t} = 0$ signal, A's belief that B is of type-0 (type-1, respectively) is $\frac{1+\epsilon}{2}$ ($\frac{1-\epsilon}{2}$, respectively) for some value of ϵ (and vice versa in the case of the $\hat{t} = 1$ signal with the same ϵ). We stress that B's BNE strategy is not Randomized Response—the different types of B agent do not place the same probability on sending the signal that best matches their respective types ($\hat{t} = t$).

(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 his or 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; hence B's BNE strategy in the simplest of all signaling games does not follow Randomized Response. However, if B's valuation for the coupon is 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

[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.

signals truthfully $\hat{t} = t$. Hence, an A agent who does not know B's valuation thinks of B as following a randomized strategy, where under a certain regime of parameters this randomized strategy could be interpreted by A as Randomized Response.

(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 get in return a fixed payment (i.e., $P(\perp, t) = \epsilon$ for each $t \in \{0, 1\}$).[3] 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 and in this BNE strategy B indeed plays like in the Randomized Response. This is a direct result of the opt-out payment—B's BNE strategy is such that it skews the sent type just slightly in favor of his or her true type, yet maintains A's (weak) preference to opt out over making an accusation. Note, however, that under alternative settings of the game's parameters, the strategy of B is such that at least one of the two types plays deterministically.

Conclusions and future directions appear in Section 6, where we provide a discussion of our results. Our work shows that Randomized Response does relate to the behavior of classical utility-maximizing agents under concrete privacy concerns modeled in various ways. However, this relation is far from straightforward as our three different games show. In fact, finding a model in which B's BNE strategy is indeed to play Randomized Response (the game detailed in Section 5 where A has an opt-out option) took multiple attempts. 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 his or her behavior s.t. $\hat{t} = t$ w.p. close to $\frac{1}{2}$. Note also that in all games we have the same phenomenon: when the coupon's value exceeds the concrete loss due to privacy concerns, B acts deterministically as it would in the toy game (without a privacy accusation) and sends the signal that discloses his or her type.

1.1 Related Work

The study of the intersection between mechanism design and differential privacy began with the seminal work of McSherry and Talwar [19], 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 [14], 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. [21]. The work of Xiao [28] 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. [20] (who effectively proposed a preliminary version of our coupon game called "Rye or Whole Wheat") 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 [14] 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

[3]Note that we can always shift all payments in the game by a constant so that $P(\perp, t) = 0$.

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 [2, 6] studying the effect of price differentiation based on an agent allowing the company to remember whether he or 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 an 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 [16], 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 his or her actions and how his or her actions are perceived by a third party.

The privacy games that we propose and analyze in this article fall into the class of signaling games [18], 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 [25]). 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 his or her type deterministically.

2 PRELIMINARIES

2.1 Equilibrium Concept

We model the games between A and B as Bayesian extensive-form games. A *Bayesian* game between two agents A and B is specified by their type spaces (Γ_A, Γ_B), a prior distribution Π 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, $u_i : \Gamma_A \times \Gamma_B \times C_A \times C_B \to \mathbb{R}$, $i \in \{A, B\}$. A *mixed* or *randomized* strategy of agent i maps a type of agent i to a distribution over his or her available actions, i.e., $\sigma_i : \Gamma_i \to \Delta(C_i)$, where $\Delta(C_i)$ is the probability simplex over C_i. When σ_i deterministically maps a type to an action, it is called a *pure strategy*. The BNE of the two-player game is defined as follows.

Definition 2.1. 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'.

In other words, a strategy profile (σ_A, σ_B) is a BNE if both agents maximize their expected utility by playing σ_i in responding to the other player's strategy σ_{-i}; i.e., they both play *best response*. In our particular case, where the game is sequential (B moves first, then A takes an action based on B's action), we have that the set of possible actions of B is simple to describe: $C_B = \{\texttt{send } \hat{t} = 0, \texttt{send } \hat{t} = 1\}$, whereas the set of possible actions of A is *dependent on the action of* B, and thus, A's set of possible actions is composed of 2-tuples: each element of C_A is of the form $\langle \tilde{t}_0, \tilde{t}_1 \rangle$, implying A plays $\tilde{t}_0$ in response to B's signal $\hat{t} = 0$ and $\tilde{t}_1$ in response to $\hat{t} = 1$. It follows that the BNE strategy of B maps his or her type to an action $\sigma_B^*(t) \mapsto \Delta(C_B)$; and we often use the notation $p^* = \mathbf{Pr}[\sigma_B^*(0) =$

0] (thus an agent of type $t = 0$ sends the signal $\hat{t} = 0$ w.p. p^* and the signal $\hat{t} = 1$ w.p. $1 - p^*$) and $q^* = \mathbf{Pr}[\sigma_B^*(1) = 1]$ (so an agent of type $t = 1$ sends the signal $\hat{t} = 0$ w.p. $1 - q^*$ and the signal $\hat{t} = 1$ w.p. q^*).

As mentioned in Section 1.1, our games between A and B belong to the class of signaling games: B is the *sender* and A is the *receiver*. In signaling games, an important subclass of dynamic/sequential games, a commonly used refinement of BNE is an equilibrium concept called *Perfect Bayesian Equilibrium*. PBEs are composed of the strategies for each player in each stage of the game, concatenated with his or her posterior belief after witnessing the history at each stage. (These beliefs are the posterior distributions on each type and are consistent on the witnessed history and the equilibrium strategies of all other players.) For brevity, we avoid formally defining the more subtle concept of PBE as the refinement doesn't provide additional insights for our problem and alleviates us of the need to explicitly introduce the description of A's posterior belief based on the observed signal $\hat{t}$. It is, however, evident that all of the BNEs considered in our article can be "extended" to PBEs (by appropriately defining the beliefs of agent A about agent B's type post witnessing the signal $\hat{t}$) as A's BNE strategy σ_A^* must be such that the actions taken in response to B sending $\hat{t}$ are best response w.r.t to the resulting posterior distribution: $\left(\frac{D_0 p^*}{D_0 p^* + D_1(1-q^*)}, \frac{D_1(1-q^*)}{D_0 p^* + D_1(1-q^*)}\right)$ after witnessing $\hat{t} = 0$, and $\left(\frac{D_0(1-p^*)}{D_0(1-p^*) + D_1 q^*}, \frac{D_1 q^*}{D_0(1-p^*) + D_1 q^*}\right)$ after witnessing $\hat{t} = 1$.[4] For signaling games, the terms *separating equilibrium* and *pooling equilibrium* are often used to characterize when signaling is fully informative. At a separating equilibrium, a sender's strategy allows the receiver to deterministically infer his or her private type, while at a pooling equilibrium multiple types of senders may take the same action, preventing the receiver from gaining any information about his or her type. In contrast, our work is centered on finding a semipooling equilibrium that allows the receiver to infer limited information as to the sender's true type. We refer the interested reader to [12] (Ch. 8.2) for more information regarding PBE in signaling games.

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 2.2 ([9]). An algorithm ALG that maps inputs into some range $\mathcal{R}$ satisfies *ϵ-differential privacy* if for all pairs of neighboring inputs $\mathcal{I}, \mathcal{I}'$ and all subsets $\mathcal{S} \subset \mathcal{R}$ it holds that $\mathbf{Pr}[\text{ALG}(\mathcal{I}) \in \mathcal{S}] \leq e^{\epsilon}\mathbf{Pr}[\text{ALG}(\mathcal{I}') \in \mathcal{S}]$.

One of the simplest algorithms that achieves ϵ-differential privacy is called *Randomized Response* [10, 17], which dates back to the 1960s [26]. This algorithm is best illustrated over a binary input, where each individual is represented by a single binary bit (therefore, a neighboring instance is one in which a single 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. $\mathbf{Pr}[\hat{b}_i = b_i] = \frac{1+\gamma}{2}$ for some $\gamma \in [0, 1)$. It follows from the definition of the algorithm that it satisfies ϵ-differential privacy for $\epsilon = \ln(\frac{1+\gamma}{1-\gamma})$. Randomized Response is sometimes presented as a distributed algorithm, where each individual generates his or her respective $\hat{b}_i$ locally, which he or she reports 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 $\hat{b}_i$.

[4] Under the assumption that $p^* + 1 - q^* > 0$ and $1 - p^* + q^* > 0$; namely, that both signals are sent with nonzero probability by B at equilibrium.

In this work, we define certain games and analyze the behavior of the two types of B agents 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[5]

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

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 his or her BNE strategy σ_B^* satisfies $\mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1] = p$ for some $p \in [1/2, 1)$.

2.2.1 Privacy-Aware Agents. The notion of privacy-aware agents has been developed through a series of works [4, 14, 20, 28]. 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 [14]—i.e., the agent's privacy loss when interacting with a mechanism that satisfies ϵ-differential privacy is exactly $v \cdot \epsilon$ for some $v > 0$. (We also briefly discuss later how to extend this result to other models of u_B^{priv}.)

Recall our (simplest) toy game, in which all that happens is that B asks for 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 (and valuation of 0 a coupon of type $1-t$). Therefore, in this game, $u_B^{out} = \rho_t \mathbb{1}_{[\hat{t}=t]}$. We think of σ_B^*, B's utility-maximizing strategy 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 his or 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 2.3. *Fix $v > 0$, and assume ρ_0, ρ_1 satisfy the following two conditions. (i) There exists a constant $\alpha > 0$ such that $\min\{D_0\rho_0, D_1\rho_1\} \geq \alpha(D_0\rho_0 + D_1\rho_1)$ ($\geq \alpha \max\{D_0\rho_0, D_1\rho_1\}$, i.e., no one coupon dominates the value of the other), and (ii) ρ_0, ρ_1 are sufficiently large. Let B be a privacy-aware agent whose privacy loss is given by $v \ln(X_{\text{game}})$. Then, the unique strategy σ_B^* that maximizes B's utility is a Randomized Response strategy. That is, $\mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1] = p^*$ for some $p^* \in (\frac{1}{2}, 1)$.*

Proof. Recall the type of B is chosen randomly to be 0 w.p. D_0 and 1 w.p. D_1. Given a strategy σ_B for B, we denote $p = \mathbf{Pr}[\sigma_B(0) = 0]$ and $q = \mathbf{Pr}[\sigma_B(1) = 1]$ (so $\mathbf{Pr}[\sigma_B(0) = 1] = 1 - p$ and $\mathbf{Pr}[\sigma_B(1) = 0] = 1 - q$). Therefore,

$$X_{\text{game}}(\sigma_B) = X_{\text{game}}(p, q) = \max\left\{ \frac{p}{1-q}, \frac{1-q}{p}, \frac{q}{1-p}, \frac{1-p}{q} \right\}. \tag{1}$$

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

Note that $X_{\text{game}}(p,q) \geq 1$ with equality iff $p = 1-q$ (which means $\sigma_B(t)$ is independent of t and B reveals no information about his or her type). And so B aims to maximizes the following utility function: $u_B = D_0\rho_0 p + D_1\rho_1 q - v\ln(X_{\text{game}})$. When the strategy that optimizes B's utility, denoted (p^*,q^*), satisfies $p^* = q^* \in (\frac{1}{2},1)$, we say that B plays using Randomized Response.

First, observe that if $p+q<1$, then $X_{\text{game}} > 1$ and the utility of B is $D_0\rho_0 p + D_1\rho_1 q - v\ln(X_{\text{game}}) < \max\{D_0\rho_0, D_1\rho_1\}$, so B can always improve the utility by playing either $(p,q)=(1,0)$ or $(p,q)=(0,1)$. The same argument holds for any (p,q) where $p+q=1$ and both are not integral. (If $D_0\rho_0 = D_1\rho_1$, then the agent is indifferent between any (p,q) satisfying $p=1-q$.) Second, observe that the utility-maximizing strategy cannot be (p,q) with $p=1$ and $q>0$ (or with $q=1$ and $p>0$) because in that case the privacy loss is infinite. Therefore, we deduce the following. If there exists a strategy (p,q) s.t. $p>1-q$ and $p,q\in(0,1)$ whose utility is strictly greater than $\max\{D_0\rho_0, D_1\rho_1\}$, then it is a utility-maximizing strategy. (Otherwise, one of the two strategies $(0,1)$ or $(1,0)$ maximizes B's utility.)

We continue assuming wlog that $D_0\rho_0 \geq D_1\rho_1$. Note that in this case the utility-maximizing strategy (p^*,q^*) must satisfy $p^* \geq q^*$, since otherwise the strategy (q^*,p^*) yields greater utility without changing X_{game}. This implies that we seek to maximize B's utility over the domain $\{(p,q) : p+q\geq 1, p\geq q\}$, where $X_{\text{game}} = \frac{q}{1-p}$ and B's utility function is the function $f(p,q) = D_0\rho_0 p + D_1\rho_1 q - v\ln(\frac{q}{1-p})$. Observe that f is a convex function of q, and so the maximum must be obtained when $q=p$ or when $q=1-p$. As we rule out the latter option, we have that $p=q$.

We now ask whether there exists a value $x^* > \frac{1}{2}$ such that B gains more than $D_0\rho_0$ by playing $(p,q)=(x^*,x^*)$. That is, we ask

$$\exists ? x > \tfrac{1}{2} \text{ s.t. } (D_0\rho_0 + D_1\rho_1)x - v\ln\left(\frac{x}{1-x}\right) > D_0\rho_0.$$

Denoting $Y = D_0\rho_0 + D_1\rho_1$, we now use the assumption that some constant α exists s.t. $D_1\rho_1 \geq \alpha Y$ to deduce that $D_0\rho_0 \leq (1-\alpha)Y$. This simplifies our question to whether $x>\frac{1}{2}$ does exist s.t. $Y\cdot x - v\ln(\frac{x}{1-x}) > (1-\alpha)Y$. This last function is maximized for x^* satisfying $x^*(1-x^*) = \frac{v}{Y}$, i.e., $x^* = \frac{1}{2}(1+\sqrt{1-\frac{4v}{Y}})$. Therefore, B will prefer playing this randomized strategy if

$$u_B(x^*,x^*) = \tfrac{1}{2}Y\left(1+\sqrt{1-\tfrac{4v}{Y}}\right) - v\ln\left(\frac{1+\sqrt{1-\frac{4v}{Y}}}{1-\sqrt{1-\frac{4v}{Y}}}\right) > (1-\alpha)Y \geq D_0\rho_0.$$

Since $\lim_{Y\to\infty}\frac{u_B(x^*,x^*)}{Y} = 1$, then for a large enough value of Y, the above inequality holds. □

A general privacy valuation. As an immediate corollary of the proof, consider any alternative definition of a privacy-aware agent in which the privacy valuation u_B^{priv} (1) depends only on the strategy σ_B, (2) is nonnegative, (3) is upper bounded by $v\ln(X_{\text{game}})$ for some $v>0$, and (4) $u_B^{priv} = \infty$ whenever $X_{\text{game}} = \infty$. We argue that the utility-maximizing strategy of such an agent is also randomized. (Observe that we no longer guarantee that B's optimal strategy σ_B^* satisfies $\mathbf{Pr}[\sigma_B^*(0)=0] = \mathbf{Pr}[\sigma_B^*(1)=1]$.)

To see that, observe that whenever $p=1-q$, we have that $X_{\text{game}}=0$ so the privacy loss of an agent is 0. Therefore, playing either $(p,q)=(1,0)$ or $(0,1)$, the agent can guarantee a utility of $\max\{D_0\rho_0, D_1\rho_1\}$. In contrast, should the agent play any (p,q) with $p<1-q$, then his or her utility is upper bounded by $D_0\rho_0 p + D_1\rho_1 q \leq \max\{D_0\rho_0, D_1\rho_1\}$, because the privacy loss is nonnegative. Therefore, the agent prefers playing $(p,q)=(1,0)$ or $(0,1)$ to any (p,q) with $p<1-q$. Second, since we assume infinite privacy loss whenever $X_{\text{game}}=\infty$, then B's utility-maximizing strategy cannot satisfy that $p=1$ and $q>0$ (or vice versa). Lastly, the proof of Theorem 2.3 gives a strategy

(p, q) with $p > 1 - q$ where the lower bound on B's utility is greater than $\max\{D_0\rho_0, D_1\rho_1\}$. It follows that B strictly prefers playing some strategy (p, q) with $p, q \in (0, 1)$ over playing $(p, q) = (1, 0)$ or $(p, q) = (0, 1)$.

2.2.2 The Two Types of B Agents as Different Players. The above analysis assumed B is an agent playing this coupon game, decides on a strategy before the realization of his or her type, and sticks to that strategy even after his or her type is revealed. It is possible, though, to think of the two types of B agents as two different agents ex-post—after each agent is revealed his or her own type. As we show, the analysis in this case is slightly different. Observe that in this case we discuss a straightforward Nash equilibrium, as both agents know their respective types. In the following, we continue using our notation from earlier, where $\mathbf{Pr}[\sigma(i) = \hat{t}]$ denotes the probability a B agent of type $t = i$ sends the signal $\hat{t}$ according to strategy σ.

Theorem 2.4. *Consider the two-player game where player $i \in \{0, 1\}$ is a B agent of type $t = i$. Assume $\rho_0 = \rho_1 = \rho$ and that ρ is sufficiently large. Let $z^* \in (1 - \frac{v}{\rho}, 1)$ be the number that satisfies $2z^* - 1 = \frac{v}{\rho} \ln(\frac{z^*}{1-z^*})$. Then any NE of the game falls into one of three categories:*

- $\mathbf{Pr}[\sigma(0) = 0] = \mathbf{Pr}[\sigma(1) = 0] = z$ *for some* $z \in [z^*, 1]$. *(Both agents take the same strategy and send the signal $\hat{t} = 0$ with high probability z.)*
- $\mathbf{Pr}[\sigma(0) = 1] = \mathbf{Pr}[\sigma(1) = 1] = z$ *for some* $z \in [z^*, 1]$. *(Both agents take the same strategy and send the signal $\hat{t} = 1$ with high probability z.)*
- $\mathbf{Pr}[\sigma(0) = 0] = \mathbf{Pr}[\sigma(1) = 1] = z$ *for some* $z \in [1 - \frac{v}{\rho}, z^*]$. *(Both agents play Randomized Response and report truthfully $\hat{t} = t$ with the same probability z.)*

Proof. We continue using the same notation from Theorem 2.3: $p = \mathbf{Pr}[\sigma(0) = 0]$ and $q = \mathbf{Pr}[\sigma(1) = 1]$, and so $X_{\text{game}} = X_{\text{game}}(p, q)$ as denoted in Equation (1). In particular, when $p + q \geq 1$, it holds that $X_{\text{game}}(p, q) = \frac{p}{1-q}$ when $p \leq q$, and $X_{\text{game}}(p, q) = \frac{q}{1-p}$ when $p \geq q$.

First, observe that the utilities of both agents are symmetric: $u_{B,0}(p, q) = \rho p - v \ln(X_{\text{game}}(p, q))$ and $u_{B,1}(p, q) = \rho q - v \ln(X_{\text{game}}(p, q))$. Second, observe that if one agent plays deterministically $\mathbf{Pr}[\sigma(t) = \hat{t}] = 1$, then, unless the other type deterministically sends the same signal, $X_{\text{game}}(p, q) = \infty$, causing both agents to have utility of $-\infty$. It is therefore clear that the strategies $(p, q) = (1, 0)$ and $(p, q) = (0, 1)$ are both NEs. Second, observe that if (p, q) is an NE of the game, then it must hold that $p + q \geq 1$. Otherwise, if $p + q < 1$ and wlog $p < 1/2$, then type $t = 0$ agent can deviate to playing $1 - q$ and only increase his or her utility.

Before continuing with our analysis, we discuss the following two functions:

- For any parameter $q \in (0, 1)$, we denote $f_q(x) = \rho x - v \ln(\frac{q}{1-x})$. Since $f_q'(x) = \rho - \frac{v}{1-x}$ is a decreasing function on the interval $[0, 1)$, we have that f_q is maximized at $x = 1 - \frac{v}{\rho}$. In particular, f_q is strictly increasing on the interval $[0, 1 - \frac{v}{\rho}]$ and strictly decreasing on the interval $[1 - \frac{v}{\rho}, 1)$.
- For any parameter $q \in (0, 1)$, we denote $g_q(x) = \rho x - v \ln(\frac{x}{1-q})$. Since $g_q'(x) = \rho - \frac{v}{x}$, it is an increasing function on the interval $(0, 1]$, and then g_q is strictly decreasing on the $(0, \frac{v}{\rho})$ interval and strictly increasing on the $[\frac{v}{\rho}, 1]$ interval.

We return to our NE analysis. To find the remaining NEs of the game, we fix a certain strategy for the $t = 1$ agent, denoted $q = \mathbf{Pr}[\sigma(1) = 1]$, and see what strategy $p = \mathbf{Pr}[\sigma(0) = 0]$ is type $t = 0$ agent's best response for q. Since both agents are symmetric, our analysis also translates to a best-response analysis for type $t = 1$ agent.

Assume $q = \mathbf{Pr}[\sigma(1) = 1] < \frac{1}{2}$ for now. As we have $p + q \geq 1$, it must hold that $p \in [1 - q, 1)$. Since $p \geq 1 - q \geq \frac{1}{2} > q$, $X_{\text{game}} = \frac{q}{1-p}$, so type $t = 0$ agent's utility from playing $\mathbf{Pr}[\sigma(0) = 0] = p$ is precisely $f_q(p)$. There are two cases to consider here:

- When $q \leq \frac{v}{\rho}$: it implies that $1 - \frac{v}{\rho} \leq 1 - q \leq p < 1$, and so on the interval $[1 - q, 1)$ we have that f_q is strictly decreasing. Therefore, type $t = 0$ agent's best response to q is to play $p = 1 - q$.
- When $\frac{v}{\rho} < q < \frac{1}{2}$: it implies that the point $1 - \frac{v}{\rho}$ lies inside the interval $[1 - q, 1)$, so type $t = 0$ agent's best response to q is to play $p = 1 - \frac{v}{\rho}$.

Assume now the case $\frac{1}{2} \leq q$. Now, the type $t = 0$ agent can play a strategy p that lies either in the interval $[1 - q, q]$ or in the interval $[q, 1)$. In the former case, the utility of the type $t = 0$ agent is $g_q(p)$, and in the latter case his or her utility is $f_q(x)$. Therefore:

- When $\frac{1}{2} \leq q < 1 - \frac{v}{\rho}$: This implies $[1 - q, q] \subset (\frac{v}{\rho}, q]$, so the strategy that maximizes g_q on the interval $[1 - q, q]$ is to play $p = q$, where type $t = 0$ agent gains $g_q(q) = \rho q - v \ln(\frac{q}{1-q}) = f_q(q)$. Also, it holds that the point $1 - \frac{v}{\rho}$ lies inside the interval $(q, 1)$, so the strategy that maximizes f_q on this interval is $p = 1 - \frac{v}{\rho}$, where he or she gains $f_q(1 - \frac{v}{\rho}) > f_q(q)$. It follows that type $t = 0$ agent's best response to q is to set $p = 1 - \frac{v}{\rho}$.
- When $1 - \frac{v}{\rho} \leq q < 1$: This implies that in order to maximize the function g_q on $[1 - q, q]$, type $t = 0$ agent considers the two strategies $p = 1 - q$ and $p = q$; and in order to maximize the function f_q on $[q, 1)$, type $t = 0$ agent considers solely the strategy $p = q$. (Note that $f_q(q) = g_q(q)$.) Therefore, type $t = 0$ agent's best response is $p = q$ whenever $g_q(q) = \rho q - v \ln(\frac{q}{1-q}) \geq \rho(1 - q) = g_q(1 - q)$; and his or her best response is $p = 1 - q$ whenever $g_q(q) \leq g_q(1 - q)$. Observe that $g_q(q) = g_q(1 - q)$ precisely for $q = z^*$. So for $q \in [1 - \frac{v}{\rho}, z^*]$, the best response is to set $p = q$, and for $q \in [z^*, 1)$, the best response is to set $p = 1 - q$.

Because of the symmetry between the two types of agents, it follows that the NEs of the game are $(1 - q, q)$ for any $q \in [z^*, 1]$; $(1 - q, q)$ for any $q \in [0, 1 - z^*]$; and (q, q) for any $q \in [1 - \frac{v}{\rho}, z^*]$. (Note, for $q \in (1 - z^*, \frac{v}{\rho})$, type $t = 0$ agent's best response is to play $p = 1 - q$, but then type $t = 1$ agent's best response is to deviate to $q = 1 - \frac{v}{\rho}$.) □

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. (1) Proper scoring rules assign profit to A based on the accuracy of his or her belief, so A has incentives to improve his or her prior belief on B's type. (2) As we show, in this model it is possible to quantify B's tradeoff between an ϵ-change in the belief and the cost that B pays A. In that aspect, this model gives a clear quantifiable tradeoff 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 [13] (yet in a very different capacity).

Proper scoring rules (see surveys [15, 27]) 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 him or 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 $x = \mathbf{Pr}[X = 1]$. The 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 Appendix A.

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. He or she has utility that is directly linked to his or her posterior belief on B's type, and A reports his or 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 straightforward (and more boring) game where B does nothing, A merely reports x—his or 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 $F_{D_1}(x) \stackrel{\text{def}}{=} D_0 f_0(x) + D_1 f_1(x)$ in expectation. Since (f_0, f_1) is a proper scoring rule, A maximizes his or her expected payment by reporting $x = D_1$. So, in this game A gets paid $g(D_1) \stackrel{\text{def}}{=} f_{D_1}(D_1)$ 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 one of two possible coupons, knowing that agents of type t prefer a coupon of type t. And so, B chooses what type to report to A, who subsequently 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 3.1. *Consider the coupon game with payments in the form of a symmetric proper scoring rule where both types of agents have the same value for the coupon ρ $(= \rho_0 = \rho_1)$. Assume also the following about ρ, the value of the coupon:*

$$f_1(D_0) - f_1(D_1) < \rho < f_1(1) - f_1(0) \; (= f_0(0) - f_0(1)).$$

Then the unique BNE strategy of B in this game, denoted σ_B^*, satisfies that $\mathbf{Pr}[t=0 \mid \sigma_B^*(t)=0] = \mathbf{Pr}[t=1 \mid \sigma_B^*(t)=1]$.

Prior to presenting the proof of Theorem 3.1 (in Section 3.2), we wish to make a few observations and comments.

Comparison with Randomized Response. 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 strategy is different from the BNE strategy σ_B^* given in Theorem 3.1 when $\mathbf{Pr}[t=0] \neq \mathbf{Pr}[t=1]$ (i.e., $D_0 \neq D_1$). Yet, in this game, a rational agent B plays s.t. A's posterior on B's type is symmetric. Specifically, the proof of Theorem 3.1 is that the BNE strategy of B satisfies

$$\frac{D_0 p^*}{D_0 p^* + D_1(1-q^*)} = \frac{D_1 q^*}{D_0(1-p^*) + D_1 q^*} \;\Rightarrow\; D_0^2 p^*(1-p^*) = D_1^2 q^*(1-q^*), \tag{2}$$

and so, unless $D_0 = D_1$, we have that $p^* \neq q^*$.

The viewpoint of A. We also comment about A's payment. The proof of Theorem 3.1 shows that when $\hat{t} = 0$, A gets an expected payment of $g(y_0)$, and when $\hat{t} = 1$, A gets $g(y_1)$, where $y_0 = \frac{D_0 p^*}{D_0 p^* + D_1(1-q^*)}$ and $y_1 = \frac{D_1 q^*}{D_0(1-p^*) + D_1 q^*}$ (and p^* and q^* are set such that $y_0 = 1 - y_1$). Now, since the scoring rule is symmetric, we have that A gets the same payment regardless of the signal, so A's payment is $g(y_1)$. Recall that y_1 is the point where $\rho = g'(y_1)$.

So, is this game worthwhile for A? Suppose A could choose between either this coupon game or the "benchmark game" mentioned in the beginning of this section, in which A guesses B's type based solely on the known prior distribution (without viewing any signal from B). Recall, in the benchmark game, A gets an expected profit of $g(D_1) = g(D_0)$. Thus, we ask, when is it the case that $g(y_1) > g(D_0)$?[6]

Recall that g is a convex function that is minimized at $x = \frac{1}{2}$. Therefore, $g(y_1) > g(D_0)$ iff $\frac{1}{2} \leq D_0 < y_1$, which also implies $g'(D_0) < g'(y_1) = \rho$. In other words, A gains more money in the coupon game than in the benchmark game only if A offers a coupon of high-value—a coupon whose value exceeds $g'(D_0)$.

The case with $\rho_0 \neq \rho_1$. We briefly discuss the case where ρ_0 and ρ_1 are not equal. First of all, observe that now there could be situations in which the BNE is of the form $(1, q^*)$ with a nonintegral q^*, or the symmetric $(p^*, 1)$—when the coupon's value of the respective type is large enough to compensate for A knowing for certain B's type. However, when the coupon's value isn't so large, both types of B agents play a randomized strategy and we can show that the resulting posterior distributions (given by y_0 and y_1 as detailed above) satisfy

$$\frac{\rho_1}{\rho_0 + \rho_1} f_0(y_0) + \frac{\rho_0}{\rho_0 + \rho_1} f_1(y_0) = \frac{\rho_1}{\rho_0 + \rho_1} f_0(y_1) + \frac{\rho_0}{\rho_0 + \rho_1} f_1(y_1).$$

In other words, setting $\mu = \frac{\rho_0}{\rho_0 + \rho_1}$, we have $F_\mu(y_0) = F_\mu(y_1)$. Sadly, it is no longer the case that $y_1 = 1 - y_0$.

3.2 Proof of Theorem 3.1

We now present the proof of Theorem 3.1.

Proof. We first analyze both agents' utilities and strategies. The utility of A is solely based on the payments of the proper scoring rule: $E_{t \leftarrow \{D_0, D_1\}}[f_t(x)]$. A has to decide on two potential reports: x_0 and x_1, where for $b \in \{0, 1\}$, x_b represents A's belief about $\mathbf{Pr}[t = 1 \mid \hat{t} = b]$. Therefore, a strategy σ_A of A maps a signal $\hat{t}$ into a report. The utility of B has two components—B gains a certain amount of utility ρ_t from reporting A the true type but then has to pay A his or her scoring rule payments. Therefore, a strategy σ_B maps each of B's types to a signal. Given a strategy σ_B, we use the following notation:

$$p = \mathbf{Pr}[\sigma_B(0) = 0], \quad q = \mathbf{Pr}[\sigma_B(1) = 1].$$

This way, B's utility function takes the form

$$\begin{aligned} u_B &= D_0 u_{B,0} + D_1 u_{B,1}, \\ \text{where} \quad u_{B,0} &= p\,(\rho - f_0(x_0)) + (1-p)\,(-f_0(x_1)) \\ u_{B,1} &= q\,(\rho - f_1(x_1)) + (1-q)\,(-f_1(x_0))\,. \end{aligned}$$

[6] Again, we assume the coupon costs A nothing. So A merely chooses whether to (1) not offer B the coupon and get an expected reward of $g(D_1)$ or (2) offer B the coupon and obtain a reward of $g(y_1)$.

When A sees the signal $\hat{t}$, the probability over B's type is given by Bayes Rule:

$$y_0 = y_0(p,q) \stackrel{\text{def}}{=} \mathbf{Pr}[t=1 \mid \hat{t}=0] = \frac{D_1(1-q)}{D_0 p + D_1(1-q)} = \frac{1}{1+\frac{D_0 p}{D_1(1-q)}} \tag{3}$$

$$y_1 = y_1(p,q) \stackrel{\text{def}}{=} \mathbf{Pr}[t=1 \mid \hat{t}=1] = \frac{D_1 q}{D_0(1-p) + D_1 q} = \frac{1}{1+\frac{D_0(1-p)}{D_1 q}}, \tag{4}$$

and since A's payments come from a proper scoring rule, it follows that A reports $x_0 = \sigma_A(0) = y_0$ and $x_1 = \sigma_A(1) = y_1$. In other words, given that B's BNE strategy is (p^*, q^*), then A plays best response of $x_0^* = y_0(p^*, q^*), x_1^* = y_1(p^*, q^*)$.

We now turn to analyze B's utility. Denote the strategy that A plays as x_0 and x_1. Then agent B decides on p and q that maximize the utility function

$$u_B = D_0 \cdot (p(\rho - f_0(x_0)) - (1-p)f_0(x_1)) + D_1 \cdot (q(\rho - f_1(x_1)) - (1-q)f_1(x_0)).$$

It is simple to characterize B's best response to A's strategy of (x_0, x_1).

$$\begin{aligned}
&\text{If } \rho > f_0(x_0) - f_0(x_1), \text{ then } p = 1.\\
&\text{If } \rho < f_0(x_0) - f_0(x_1), \text{ then } p = 0.\\
&\text{If } \rho = f_0(x_0) - f_0(x_1), \text{ then } B \text{ may play any } p \in [0,1].\\
&\text{If } \rho > f_1(x_1) - f_1(x_0), \text{ then } q = 1.\\
&\text{If } \rho < f_1(x_1) - f_1(x_0), \text{ then } q = 0.\\
&\text{If } \rho = f_1(x_1) - f_1(x_0), \text{ then } B \text{ may play any } q \in [0,1].
\end{aligned} \tag{5}$$

We now wish to characterize the game's BNEs. First, we claim that in a BNE, with B playing $\sigma_B^* = (p^*, q^*)$, it cannot be that $p^* < 1 - q^*$. This follows from the fact that $y_0(p,q) > y_1(p,q) \Leftrightarrow p < 1 - q$. It means that A's best response to such (p^*, q^*) is to answer some (x_0, x_1) s.t. $x_0 > x_1$. But since f_0 is a decreasing function, f_1 is an increasing function, and $\rho > 0$, then B's best response to such (x_0, x_1) is to deviate to $(1,1)$. Similarly, should (p^*, q^*) be such that $p^* = 1 - q^*$ *and both* $p^*, q^* \in (0,1)$, then A's best response (x_0, x_1) is $(\frac{1}{2}, \frac{1}{2})$, which implies again that B prefers to deviate to $(1,1)$. It follows that, with the exception of $(1,0)$ and $(0,1)$, any BNE strategy of B satisfies $p^* > 1 - q^*$, and so any BNE strategy of A satisfies $x_0 < x_1$.

Before continuing with the proof, we would like to make two observations, which we will repeatedly use. Let X be a uniform Bernoulli random variable. We examine the expected payment to an expert reporting a belief of z as to the probability of the event $X = 1$, which we denote as $F_{1/2}(z) = \frac{1}{2}(f_0(z) + f_1(z))$. The function $F_{1/2}$ is a concave function with a unique maximum at $z = \frac{1}{2}$, and it is strictly increasing on the $[0, \frac{1}{2}]$ interval and strictly decreasing on $[\frac{1}{2}, 1]$ interval. Therefore, for any a there exists at most two distinct preimages $z_1 \le \frac{1}{2} \le z_2$ satisfying $F_{1/2}(z_1) = F_{1/2}(z_2) = a$. Recall that we assume (f_0, f_1) is a symmetric proper scoring rule (so $f_1(z) = f_0(1-z)$ for any $z \in [0,1]$). So our first observation is: for any z_1, z_2 satisfying $F_{1/2}(z_1) = F_{1/2}(z_2)$ and $z_2 > z_1$, we have that $z_2 = 1 - z_1$ with $z_1 \in [0, 1/2)$ and $z_2 \in (1/2, 1]$. Using again the fact that (f_0, f_1) is a symmetric proper scoring rule and the fact that $F_{1/2}$ is maximized at $z = \frac{1}{2}$, we make our second observation: for any z, z' satisfying $F_{1/2}(z) \ge F_{1/2}(z')$ it must hold that $|z - \frac{1}{2}| \le |z' - \frac{1}{2}|$, which implies that $z \in [z', 1 - z']$ if $z' \le 1/2$.

We now return to the proof of the theorem using case analysis as to the potential BNE strategies of B. We will rely also on our assumption that $D_0 \ge D_1$.

(i) $(p^*, q^*) = (1,1)$, i.e., B always plays $\hat{t} = t$. This means that A sets $x_0 = 0$ and $x_1 = 1$ (i.e., A always predicts $t = b$ given the signal $\hat{t} = b$).

(∗) We deduce that if $\rho \geq f_0(0) - f_0(1)$ and $\rho \geq f_1(1) - f_1(0)$, then the game has a BNE of

$$(x_0^*, x_1^*) = (0, 1), \quad (p^*, q^*) = (1, 1).$$

We comment that since (f_0, f_1) is a symmetric proper scoring rule, then we have that $f_0(0) - f_0(1) = f_1(1) - f_1(0)$.

(ii) $(p^*, q^*) = (1, 0)$, i.e., B only sends the $\hat{t} = 0$ signal. So when A sees the $\hat{t} = 0$ signal, he or she sets $x_0 = D_1$ just as in the benchmark yet. But A is indifferent as to the choice of x_1 since the $\hat{t} = 1$ signal is never sent. In order for this to be a BNE, it must hold that $f_1(x_1) - f_1(D_1) \geq \rho \geq f_0(D_1) - f_0(x_1)$ so that both types of B agents would keep sending the $\hat{t} = 0$ signal. So x_1 satisfies that $F_{1/2}(x_1) = \frac{1}{2}(f_0(x_1) + f_1(x_1)) \geq F_{1/2}(D_1) = \frac{1}{2}(f_0(D_1) + f_1(D_1))$. Based on our second observation, we have that $x_1 \in [D_1, D_0]$.

(∗) We deduce that if the parameters of the game are set such that there exists $v \in [D_1, D_0]$ satisfying both $f_0(v) \geq f_0(D_1) - \rho$ and $f_1(v) \geq f_1(D_1) + \rho$, then the game has a BNE of

$$(x_0^*, x_1^*) = (D_1, v), \quad (p^*, q^*) = (1, 0).$$

As f_1 is an increasing function, it must hold that $\rho \leq f_1(D_0) - f_1(D_1)$. In other words, when $\rho > f_1(D_0) - f_1(D_1)$, this cannot be a BNE.

(iii) $(p^*, q^*) = (0, 1)$. This means that B only sends the $\hat{t} = 1$ signal. So now A sets $x_1 = D_1$ but A is indifferent regarding the value of x_0. In order for B not to deviate from $(0, 1)$, x_0 should satisfy both $\rho \leq f_0(x_0) - f_0(D_1)$ and $\rho \geq f_1(D_1) - f_1(x_0)$. This implies that $F_{1/2}(x_0) \geq F_{1/2}(D_1)$ and our second observation gives that $x_0 \in [D_1, D_0]$. But observe that $f_0(x_0) \geq \rho - f_0(D_1) > f_0(D_1)$. This contradicts the fact that f_0 is a strictly decreasing function.

(iv) $p^* = 1$ while $q^* \in (0, 1)$. This means A sets $x_1 = 1$ (because only type 1 agents can send $\hat{t} = 1$), while setting $x_0 = y_0(p^*, q^*) > 0$. To keep B from deviating, x_0 should satisfy that $\rho \geq f_0(x_0) - f_0(1)$ and $\rho = f_1(1) - f_1(x_0)$. Therefore, $F_{1/2}(1) \geq F_{1/2}(x_0)$, so our observation yields the contradiction $1 \in [x_0, 1 - x_0]$.

(v) $q^* = 1$ while $p^* \in (0, 1)$. This case is symmetric to the previous case, and we get a similar contradiction using $F_{1/2}(0) \geq F_{1/2}(x_1)$.

(vi) $p^*, q^* \in (0, 1)$ with $p^* > 1 - q^*$. We know that A's best response is setting $x_0^* = y_0(p^*, q^*)$ and $x_1^* = y_1(p^*, q^*)$ and we have already shown that $y_0 < y_1$. In order for B to play best response against (y_0, y_1), we must have that $\rho = f_0(y_0) - f_0(y_1) = f_1(y_1) - f_1(y_0)$ so $F_{1/2}(y_0) = F_{1/2}(y_1)$. Based on our first observation from before, we have that $y_1 = 1 - y_0$. In other words, B picks p^* and q^* s.t. the signals $\hat{t} = 0$ and $\hat{t} = 1$ are symmetric:

$$\begin{aligned}\mathbf{Pr}[t = 1 \mid \hat{t} = 1] &= y_1 = 1 - y_0 \\ &= 1 - \mathbf{Pr}[t = 1 \mid \hat{t} = 0] = \mathbf{Pr}[t = 0 \mid \hat{t} = 0],\end{aligned}$$

so regardless of the value of b, the expression $\mathbf{Pr}[t = \hat{t} \mid \hat{t} = b]$ is the same.

Observe that we have $\rho = f_0(y_0) - f_0(y_1) = f_0(y_0) - f_0(1 - y_0) = -g'(y_0)$ or $\rho = g'(y_1)$. (Recall, (f_0, f_1) are derived using a convex function g as detailed in Section A.1.) In other words, B sets (p^*, q^*) by first finding $y_1 \in (\frac{1}{2}, 1]$ s.t. $g'(y_1) = \rho$, then finding (p^*, q^*) that satisfy Equation (2) and yield y_1. Formally, B finds (p^*, q^*) that satisfy

$$\rho = g'\left(\frac{D_1 q^*}{D_0(1 - p^*) + D_1 q^*}\right) = -g'\left(\frac{D_1(1 - q^*)}{D_0 p^* + D_1(1 - q^*)}\right). \tag{6}$$

Recall that g is convex and $g'' > 0$ on the $[0, 1]$ interval. This implies that as ρ increases, the point $y_1(p^*, q^*)$ gets further away from $\frac{1}{2}$ and closer to 1. □

Modifications of the proof for the case where $\rho_0 \neq \rho_1$. We briefly discuss the case where ρ_0 and ρ_1 are not equal. First of all, observe that now there could be situations in which the BNE is of the form $(1, q^*)$ with a nonintegral q^* or the symmetric $(p^*, 1)$. This is because the previous contradiction that held in these cases no longer holds. More interestingly, for the BNE we get that (y_0, y_1) and (p^*, q^*) still satify Equations (3) and (4), and also

$$\rho_0 = f_0(y_0) - f_0(y_1), \;\; \rho_1 = f_1(y_1) - f_1(y_0),$$

which, using $\rho_0\rho_1$, can be manipulated into

$$\frac{\rho_1}{\rho_0+\rho_1} f_0(y_0) + \frac{\rho_0}{\rho_0+\rho_1} f_1(y_0) = \frac{\rho_1}{\rho_0+\rho_1} f_0(y_1) + \frac{\rho_0}{\rho_0+\rho_1} f_1(y_1).$$

In other words, setting $\mu = \frac{\rho_0}{\rho_0+\rho_1}$, we have $F_\mu(y_0) = F_\mu(y_1)$. Alternatively, it is possible to subtract the two equalities and deduce

$$\tfrac{1}{2}(\rho_0 - \rho_1) = \tfrac{1}{2}(f_0(y_0) + f_1(y_0)) - \tfrac{1}{2}(f_0(y_0) + f_1(y_1)) = F_{1/2}(y_0) - F_{1/2}(y_1).$$

These two conditions (along with $y_0 < y_1$) dictate the value of y_0, y_1, and thus the values of (p^*, q^*). Sadly, it is no longer the case that $y_1 = 1 - y_0$.

In Appendix A.2 we discuss the implications of using specific scoring rules.

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 $I_{2\times2}$. This is of course the simplest of all signaling games. 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 he or she errs. We comment that in Section 5 we consider a more general matrix of payments.[7]

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 $u_B = \rho_t \mathbb{1}_{[\hat{t}=t]} - \mathbb{1}_{[\tilde{t}=t]}$.

[7] Moreover, if we consider a general 2×2-payment matrix M, the result remains qualitatively similar—some type of B agent plays deterministically (unless some specific equality happens making that type of agent indifferent to any strategy).

Theorem 4.1. *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. $\mathbf{Pr}[\sigma_B^*(t) = \hat{t}] = 1$.*

The proof of Theorem 4.1 appears in Section 4.3. Here, we make a few comments about the BNE of this game.

First, we comment that technically, it is possible to have a BNE strategy for B, which is Randomized Response. In the case where $\rho_0 = \rho_1$, 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), yet this Randomized Response strategy is not preferable to any other BNE strategy.

But, assuming $\rho_0 \neq \rho_1$, then we have that at BNE, B plays in a way where the $\hat{t} = 0$ signal leads A to play the same way he or she plays in the benchmark game (with no coupon)—to always play $\tilde{t} = 0$, because $\mathbf{Pr}[t = 0 \mid \hat{t} = 0] > \mathbf{Pr}[t = 1 \mid \hat{t} = 0]$. However, given the signal $\hat{t} = 1$, it holds that $\mathbf{Pr}[t = 0 \mid \hat{t} = 1] = \mathbf{Pr}[t = 1 \mid \hat{t} = 1]$ (since B's BNE strategy is on the ℓ_2 line). In other words, after viewing the $\hat{t} = 1$ signal, A has posterior belief on B's type of $(\frac{1}{2}, \frac{1}{2})$. Of course, if B plays the strategy $(1, 0)$, then this last statement is vacuous since the $\hat{t} = 1$ signal is never sent.

Lastly, we compare A's payment in the full game to his or her payment in the benchmark game. Recall that in the benchmark game A always accuses $\tilde{t} = 0$ and so in expectation his or her payment is D_0. As the proof of Theorem 4.1 demonstrates, the utility that A gets by playing his or her BNE strategy is $D_0 p^* + D_0(1 - p^*) = D_0$. In other words, moving from the benchmark game to this more complicated coupon game gives A no additional revenue. In fact, the only agent that gains anything is B. In the benchmark game B's utility is $-D_0$. In the coupon game, B's utility (at BNE) is $D_0(\rho_0 - 1)$ when $\rho_0 \geq \rho_1$, or $D_0(\rho_0 - 1) + D_1(\rho_1 - \rho_0)$ when $\rho_0 < \rho_1$.

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 $\mathsf{CDF}_0(x) \stackrel{\text{def}}{=} \mathbf{Pr}[\rho < x \mid t = 0]$ 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 $\{y : \mathsf{CDF}_B(y) = z\}$. Observe that now, the strategy of B maps both his or her type *and his or her valuation of the coupon* to an action $\hat{t} \in \{0,1\}$. Thus, σ_B is a randomized function from $\mathbb{R}_{\geq 0} \times \{0,1\}$ to $\{0,1\}$.

Theorem 4.2. *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 his or 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 $\mathbf{Pr}[\sigma_B^*(\rho, t) = t] = 1$, and if $\rho < y^*$, then $\mathbf{Pr}[\sigma_B^*(\rho, t) = t] = 0$.*

Observe that in continuous-value variation of the coupon game, from A's perspective—who doesn't know the realized value of the coupon—it appears that B is playing a randomized strategy. Furthermore, should the coupon valuation and the type be chosen independently (i.e., $\mathsf{PDF}_0 = \mathsf{PDF}_1$), then A views B's strategy as Randomized Response—since for a randomly chosen ρ it holds that $\mathbf{Pr}[\sigma(\rho, 0) = 0] = \mathbf{Pr}[\sigma_B(\rho, 1) = 1] = \mathbf{Pr}[\rho > y^*]$. In that case the behavior of B preserves ϵ-differential privacy for

$$\epsilon = \ln\left(\frac{D_0}{D_1}\right).$$

And so, even though it holds that at BNE both players' strategies are fairly simple, we can still argue that Randomized Response arises as a strategy of utility-maximizing agents at equilibrium. The proof of Theorem 4.2 appears in Section 4.3.

4.3 Proofs of Theorem 4.1 and 4.2

We now present the proofs of Theorems 4.1 and 4.2. We begin with the proof of Theorem 4.1.

Proof of Theorem 4.1. First, we denote the strategies of agents A and B. We denote

$$\text{For } B\text{: } p = \mathbf{Pr}[\sigma_B(0) = 0], \text{ and } q = \mathbf{Pr}[\sigma_B(1) = 1].$$
$$\text{For } A\text{: } x = \mathbf{Pr}[\sigma_A(0) = 0], \text{ and } y = \mathbf{Pr}[\sigma_A(1) = 1].$$

Using these four parameters, we analyze the utility functions of the agents of the game. We start with the utility function of A:

$$u_A = D_0 px + D_0(1-p)(1-y) + D_1 qy + D_1(1-q)(1-x).$$

This function characterizes A's best response strategy as follows. A determines $x = \mathbf{Pr}[\sigma_A(0) = 0]$ based on the relation between $D_0 p$ $(= \mathbf{Pr}[t = 0 \wedge \hat{t} = 0])$ and $D_1(1-q)$ $(= \mathbf{Pr}[t = 1 \wedge \hat{t} = 0])$—if $D_0 p$ is the larger term, then $x = 1$; if $D_1(1-q)$ is the larger term, then $x = 0$; and if both are equal, then A is free to set any $x \in [0, 1]$. Similarly, the relationship between $D_1 q = \mathbf{Pr}[t = 1 \wedge \hat{t} = 1]$ and $D_0(1-p) = \mathbf{Pr}[t = 0 \wedge \hat{t} = 1]$ determines the value of $y = \mathbf{Pr}[\sigma_A(1) = 1]$.

We therefore denote the following two lines on the $[0, 1] \times [0, 1]$ square of possible choices for p and q:

$$\ell_1 : q = 1 - \frac{D_0}{D_1} p, \text{ (i.e., } D_0 p = D_1(1-q)).$$
$$\ell_2 : q = \frac{D_0}{D_1}(1-p), \text{ (i.e., } D_0(1-p) = D_1 q).$$

These are A's "lines of indifference": when B plays $(p, q) \in \ell_1$, A is indifferent to any value of x in the range $[0, 1]$, and when B plays $(p, q) \in \ell_2$, A is indifferent between any value of y.

Observe that ℓ_1 and ℓ_2 have the same slope, and so they are parallel, and that the point $(p, q) = (1, 0)$ is above ℓ_1 yet on ℓ_2. It follows that ℓ_2 is above ℓ_1 (unless $D_0 = D_1 = \frac{1}{2}$, in which case the two lines coincide). The two lines are shown in Figure 2.

We now turn our attention to the utility functions of B. The utility of B of type $t = 0$ is

$$u_{B,0} = p \cdot (\rho_0 - x) + (1-p)(-1 + y),$$

and the utility of B of type $t = 1$ is

$$u_{B,1} = q \cdot (\rho_1 - y) + (1-q)(-1 + x),$$

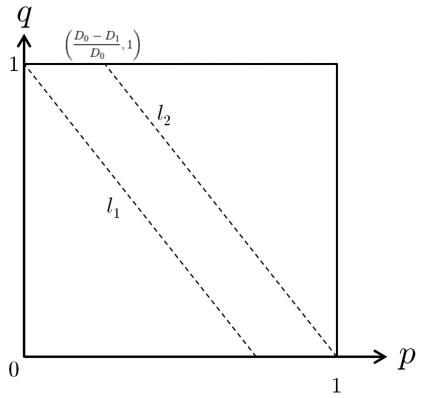

Fig. 2. The strategy space for agent B and the corresponding best response of A. When B plays a (p, q)-point above line ℓ_1, A sets $x = 1$; when B plays a (p, q)-point above line ℓ_2, A sets $y = 1$.

which means that B's best response strategies are

$$\begin{aligned}
&\text{if } \rho_0 > x + y - 1, \text{ then } p = 1;\\
&\text{if } \rho_0 < x + y - 1, \text{ then } p = 0;\\
&\text{if } \rho_0 = x + y - 1, \text{ then } B \text{ can play any } p \in [0, 1];\\
&\text{if } \rho_1 > x + y - 1, \text{ then } q = 1;\\
&\text{if } \rho_1 < x + y - 1, \text{ then } q = 0;\\
&\text{if } \rho_1 = x + y - 1, \text{ then } B \text{ can play any } q \in [0, 1].
\end{aligned}$$

Using the best response strategies of both A and B, we can analyze the game's potential BNEs. First we cover the simple case.

If $\max\{\rho_0, \rho_1\} > 1$, then at least one coupon has value strictly greater than 1 and so one of the two types of B agents strictly prefers deviating to playing deterministically. Wlog this type is the $t = 0$ type, and so in any BNE of the game we have that $\mathbf{Pr}[\sigma_B^*(0) = 0] = 1$.

The interesting case is when $\max\{\rho_0, \rho_1\} \leq 1$, and since we assume $\rho_0 \neq \rho_1$, then for some type of B agent the value of the coupon is strictly less than 1. (This intuitively makes sense—the coupon game becomes interesting only when B's value for the coupon is below the max-payment from B to A, and hence B has incentive to hide his or her true type.)

We continue with a case analysis as to the potential BNE strategies of B.

(i) Strictly above the ℓ_2 line, where $D_1 q > D_0(1 - p)$.
This means B plays s.t. $D_1\mathbf{Pr}[\sigma_B^*(1) = 1] > D_0\mathbf{Pr}[\sigma_B^*(0) = 1]$, and as a result

$$D_1\mathbf{Pr}[\sigma_B^*(1) = 0] = D_1 - D_1 q < D_0 - D_0(1 - p) = D_0\mathbf{Pr}[\sigma_B^*(0) = 0].$$

Therefore, given that A observes any signal $\hat{t} \in \{0, 1\}$, it is more likely that a B agent of type $\hat{t}$ sent that signal. So A responds to such strategy by playing deterministically $\mathbf{Pr}[\sigma_A^*(\hat{t}) = \hat{t}] = 1$ for any signal $\hat{t} \in \{0, 1\}$. As A prefers to play $x = y = 1$ and some type of B agent has coupon valuation <1, then that type deviates (so either $p = 0$ or $q = 0$), and so the BNE strategy of B cannot be above the ℓ_2 line.

(ii) Strictly below the ℓ_2 line, where $D_1 q < D_0(1 - p)$.
This means B plays s.t. $D_1\mathbf{Pr}[\sigma_B^*(1) = 1] < D_0\mathbf{Pr}[\sigma_B^*(0) = 1]$. So the $\hat{t} = 1$ signal is more likely to come from a $t = 0$ type agent, and so A's best response is to set

$(1-y) = \mathbf{Pr}[\sigma_A^*(1)=0] = 1$. We thus have that $x+y-1 \le 0$, whereas $\rho_0, \rho_1 > 0$. Hence, B deviates to playing $(p,q) = (1,1)$, and so the BNE of B cannot be below the ℓ_2 line.

(iii) On the ℓ_2 line, where $D_1 q = D_0(1-p)$.
This means B plays s.t. $D_1\mathbf{Pr}[\sigma_B^*(1)=1] = D_0\mathbf{Pr}[\sigma_B^*(0)=1]$, and as a result

$$D_1\mathbf{Pr}[\sigma_B^*(1)=0] = D_1 - D_1 q < D_0 - D_0(1-p) = D_0\mathbf{Pr}[\sigma_B^*(0)=0],$$

assuming $D_1 < D_0$ (the special case where $D_0 = D_1 = \frac{1}{2}$ will be discussed later). And so when A views the $\hat{t}=0$ signal, it is more likely that the type of B agent is $t=0$, so $x = \mathbf{Pr}[\sigma_A^*(0)=0] = 1$, whereas when A views the $\hat{t}=1$ signal, both types of B agents are as likely to send the signal, so A is indifferent as to the value of $y = \mathbf{Pr}[\sigma_A^*(1)=1]$.

Since $\rho_0 \neq \rho_1$, by setting the single parameter y, A can makes at most one of the two types of B agent indifferent, while the other type plays a pure strategy. In other words, B's BNE strategy can only be one of the two extreme points: $(p,q) = (1,0)$ or $(p,q) = (\frac{D_0-D_1}{D_1}, 1)$. The pure strategy of the nonindifferent type is determined by the relation between $\rho_0 \gtrless \rho_1$. So the possible BNEs are

$$\begin{aligned}
&\text{If } \rho_0 > \rho_1,\ \ \sigma_A^* = (1,y) \text{ with } y \in [\rho_1, \rho_0], \sigma_B^* = (1,0).\\
&\text{If } \rho_0 = \rho_1,\ \ \sigma_A^* = (1,\rho_0), \sigma_B^* = (p,q) \text{ with } (p,q) \in \ell_2.\\
&\text{If } \rho_0 < \rho_1,\ \ \sigma_A^* = (1,\rho_0), \sigma_B^* = (\tfrac{D_0-D_1}{D_0}, 1).
\end{aligned}$$

In the special case where $D_0 = D_1$ (the two lines are the same one), $(p=1, q=0)$ and $(p=0, q=1)$ are both BNE regardless of the A BNE strategy being any (x,y) satisfying with $\min\{\rho_0,\rho_1\} \le x+y-1 \le \max\{\rho_0,\rho_1\}$.

Observe that in the case with $\rho_0 = \rho_1 < 1$, in a BNE, B may play any strategy on the ℓ_2 line and A makes both types of B agents indifferent to the value of p,q by setting $y = \rho_0 = \rho_1$. Since the line $p=q$ (i.e., agent B plays Randomized Response) does intersect the ℓ_2 line at (D_0, D_0), it is possible that B plays Randomized Response (with $\epsilon = \ln(D_0/D_1)$). (And if $\rho_0 = \rho_1 = 1$, then B may play any (p,q) on ℓ_2 or above it, whereas A plays $x = y = 1$.) □

Proof of Theorem 4.2. We assume B's parameters are sampled as follows. First, we pick a type t s.t. $\mathbf{Pr}[t=1] = D_1$ and $\mathbf{Pr}[t=0] = D_0$. Then, given t, we sample $\rho \leftarrow \mathsf{PDF}_t$, where $\mathbf{Pr}[\rho \le 0] = 0$ for both types. And while A knows D_0, D_1, PDF_0, and PDF_1, A does not know B's realized type and valuation.

We apply the same notation from before, denoting a strategy σ_B of B using p and q (where $p = \mathbf{Pr}[\sigma_B(0)=0]$ and $q = \mathbf{Pr}[\sigma_B(1)=1]$), and denoting a strategy σ_A of A using x and y (where $x = \mathbf{Pr}[\sigma_A(0)=0]$ and $y = \mathbf{Pr}[\sigma_A(1)=1]$).

The utility function of B remains the same:

$$u_{B,0,\rho} = p(\rho - x) + (1-p)(-1+y), \qquad u_{B,1,\rho} = q(\rho - y) + (1-q)(-1+x).$$

So B's best response to any strategy of (x,y) of A is given by

$$\sigma_B^{br}(\rho, t) = \begin{cases} t, & \text{if } \rho > x+y-1 \\ 1-t, & \text{if } \rho < x+y-1. \end{cases}$$

We call such a strategy a *threshold strategy* characterized by a parameter T, where any agent with $\rho < T$ plays $\hat{t} = 1-t$ and any agent with $\rho > T$ plays $\hat{t} = t$.[8] Clearly, in any BNE, B follows a threshold strategy for some value of T.

[8] Since ρ is sampled from a continuous distribution, the probability of the event $\rho = T$ is 0.

Therefore, since A's BNE strategy is best response to B's BNE strategy, it suffices to consider A's best response against a threshold strategy. Given that B follows a threshold strategy with threshold T, we have that A's utility function is

$$\begin{aligned} u_A &= xD_0(1-\mathsf{CDF}_0(T)) + (1-x)D_1\mathsf{CDF}_1(T) \\ &\quad + yD_1(1-\mathsf{CDF}_1(T)) + (1-y)D_0\mathsf{CDF}_0(T) \\ &= \mathsf{CDF}_B(T) + x(D_0 - \mathsf{CDF}_B(T)) + y(D_1 - \mathsf{CDF}_B(T)), \end{aligned}$$

where we use the notation $\mathsf{CDF}_B = D_0\mathsf{CDF}_0 + D_1\mathsf{CDF}_1$. As A maximizes his or her strategy, we have that A sets $x > 0$ only if $\mathsf{CDF}_B(T) \le D_0$. Similarly, $y > 0$ only if $\mathsf{CDF}_B(T) \le D_1$. Since $D_1 \le D_0$, we only have three cases to consider.

(i) If $\mathsf{CDF}_B(T) < D_1$: In this case A's best response is to set $x = y = 1$ and B's best response to $(1,1)$ is to set the threshold parameter $T = x + y - 1 = 1$. (So every B agent with $\rho < 1$ deterministically sends the signal $\hat{t} = 1 - t$ and any B agent with $\rho > 1$ sends the signal $\hat{t} = t$.) Clearly, if it holds that $\mathsf{CDF}_B(1) < D_1$, i.e., that the probability of a random B agent to have $\rho \le 1$ is less than D_1, then we have a BNE.

(ii) If $\mathsf{CDF}_B(T) > D_1$: In this case A's best response sets $y = 0$ and $x \in [0,1]$. As B is playing best response to A's strategy, then it means that the threshold parameter is set to $T = x - 1 \le 0$. (And so all B agents have coupon valuation $\rho > 0$ and we have that all B agents deterministically send the signal $\hat{t} = t$.) But for such T we have that $\mathsf{CDF}_B(T) \le \mathsf{CDF}_B(0) = 0 < D_1$, we get an immediate contradiction.

(iii) If $\mathsf{CDF}_B(T) = D_1$: In this case A sets $x = 1$ and is indifferent to the choice of y. Observe that B's best response to A's strategy of $(1, y)$ is to set the threshold parameter to $T = y$. We have that this is indeed a BNE if $y \in \mathsf{CDF}_B^{-1}(D_1)$. Assuming uniqueness to the inverse of CDF_B, then $\sigma_A^* = (1, y^*)$ with $y^* = \mathsf{CDF}_B^{-1}(D_1)$ is A's BNE strategy, and B's BNE strategy is a threshold strategy with the threshold parameter set to $T = y^*$. We comment that in the case where $D_0 = D_1$ and A is indifferent to the choice of x as well, the BNE strategy of A is defined using any $x^*, y^* \in [0,1]$ that satisfy $x^* + y^* \in \mathsf{CDF}_B^{-1}(D_1)$. □

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 $M = \left[\begin{array}{c|c} M_{0,0} & -M_{0,1} \\ \hline -M_{1,0} & M_{1,1} \end{array}\right]$ in which 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 nonnegative. Indeed, when previously considering the identity matrix payments, we assumed that for A, realizing that B has type $t = 0$ is worth just as much as finding out that 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 $D_0 \ge D_1$) may represent having some embarrassing medical condition, while type $t = 0$ represents 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 nonembarrassing 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 revised matrix of payments just allows a brother perspective on our signaling game, yet it doesn't change the setting drastically.[9] The key change between the current setting of the game vs. the setting in Section 4 lies in A's ability to opt out and still gain something—which in turn alters the game's BNE substantially. In this new setting B knows that as long as his or her behavior is such that A (weakly) prefers opting out to making an accusation, then the accusation is worth (weakly) less then opting out in expectation. This allow B to behave in a way that skews ever so slightly his or her signal in favor of the signal that matches his or her true type. Indeed, as we prove later, in this game with an opt-out option for A, it does hold (under certain parameters) that B's BNE strategy is Randomized Response.

The benchmark game. First, prior to analyzing the full coupon game with an opt-out option, we yet again consider a much simpler setting. 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: (1) guess that B is of type 0, (2) guess that B is of type 1, and (3) guess nothing. In expectation, the outcome of option (1) is $D_0M_{0,0} - D_1M_{0,1}$ and the outcome of option (2) is $D_1M_{1,1} - D_0M_{1,0}$. 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

$$\frac{M_{0,0}}{M_{0,1}} < \frac{D_1}{D_0}\text{, and } \frac{M_{1,1}}{M_{1,0}} < \frac{D_0}{D_1}. \tag{7}$$

A direct (and repeatedly used) corollary of having A determinisitcally playing "opt out" in the benchmark game—namely, from assuming that both conditions in Equation (7) hold—is that $\frac{M_{0,0}}{M_{0,1}} < \frac{M_{1,0}}{M_{1,1}}$. This condition, $M_{0,0}M_{1,1} < M_{0,1}M_{1,0}$, can be intuitively interpreted as having a wrong "accusation" being costlier than the gain from a correct "accusation" (on average and in absolute terms).

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 5.1. *If the parameters of the game satisfy the following condition:*

$$\begin{aligned} 0 &< \rho_1 M_{1,0} - \rho_0 M_{1,1} < M_{0,1}M_{1,0} - M_{0,0}M_{1,1} \\ 0 &< \rho_0 M_{0,1} - \rho_1 M_{0,0} < M_{0,1}M_{1,0} - M_{0,0}M_{1,1}, \end{aligned} \tag{8}$$

and we also have that $D_0^2 M_{0,0}M_{1,0} = D_1^2 M_{0,1}M_{1,1}$, then the unique BNE strategy of B, denoted σ_B^*, is such that B plays Randomized Response: $\frac{1}{2} \le \mathbf{Pr}[\sigma_B^*(0) = 0] = \mathbf{Pr}[\sigma_B^*(1) = 1] < 1$.

Proving Theorem 5.1 is the goal of this section. However, before proving Theorem 5.1, we stress that the interplay between the different terms involved in Equation (8) plays a key role in determining what is the BNE of the coupon game under the opting-out option. In fact, Table 1 gives

[9] In fact, one can reiterate the analysis from Section 4 and obtain similar BNE strategies.

Table 1. The Various Conditions under Which We Characterize the BNEs of the Game

Case No.	Condition	A's Strategy (Always: $x_1 = y_0 = 0$)	B's Strategy
1	$\rho_0 \geq M_{0,0} + M_{1,0}$ and $\rho_1 \geq M_{0,1} + M_{1,1}$	$(x_0, y_1) = (1, 1)$	$(1, 1)$
2	$\rho_0 \leq M_{0,0}$ and $\frac{\rho_0}{\rho_1} \leq \frac{M_{0,0}}{M_{0,1}}$	$(x_0, y_1) = (\frac{\rho_0}{M_{0,0}}, 0)$	$P_1 = (0, 1)$
3	$0 \leq \rho_0 - M_{0,0} \leq M_{1,0}$ $\rho_1 M_{1,0} - \rho_0 M_{1,1} \geq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$	$(x_0, y_1) = (1, \frac{\rho_0 - M_{0,0}}{M_{1,0}})$	P_2
4	$\rho_1 \leq M_{1,1}$ and $\frac{\rho_0}{\rho_1} \geq \frac{M_{1,0}}{M_{1,1}}$	$(x_0, y_1) = (0, \frac{\rho_1}{M_{1,1}})$	$P_3 = (1, 0)$
5	$0 \leq \rho_1 - M_{1,1} \leq M_{0,1}$ $\rho_0 M_{0,1} - \rho_1 M_{0,0} \geq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$	$(x_0, y_1) = (\frac{\rho_1 - M_{1,1}}{M_{0,1}}, 1)$	P_4
6	$0 \leq \rho_1 M_{1,0} - \rho_0 M_{1,1} \leq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$ $0 \leq \rho_0 M_{0,1} - \rho_1 M_{0,0} \leq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$	See below	P_5

We use the notation $P_2 = (1 - \frac{D_1 M_{1,1}}{D_0 M_{1,0}}, 1)$, $P_4 = (1, 1 - \frac{D_0 M_{0,0}}{D_1 M_{0,1}})$, and $P_5 = \left(\frac{D_0 D_1 M_{0,1} M_{1,0} - D_1^2 M_{0,1} M_{1,1}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}, \frac{D_0 D_1 M_{0,1} M_{1,0} - D_0^2 M_{0,0} M_{1,0}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}\right)$. The point P_5 lies at the intersection between two specific lines, and points P_2 and P_4 are the intersection points of each of those lines with the $(q = 1)$-line and $(p = 1)$-line resp. In case 6, the strategy of A is given by $(x_0, y_1) = \frac{1}{M_{1,0}M_{0,1} - M_{0,0}M_{1,1}}(M_{1,0}\rho_1 - M_{1,1}\rho_0, M_{0,1}\rho_0 - M_{0,0}\rho_1)$.

a summary of the various relations between ρ_0, ρ_1, and the four entries of M, and the unique BNE of the game in each case. Note that the six different conditions detailed in Table 1 cover all possible settings of the game and they are also mutually exclusive (unless some inequality holds as an equality). It is plain to see that the last case in Table 1 is precisely the case discussed in Theorem 5.1. The notation in Table 1 is consistent with our notation in the analysis of the game. A strategy σ_B of agent B is denoted as (p, q) and a strategy σ_A of agent A is denoted as (x_0, x_1, y_0, y_1). Formally, we denote $p = \mathbf{Pr}[\sigma_B(0) = 0]$ and $q = \mathbf{Pr}[\sigma_B(1) = 1]$, and $x_b = \mathbf{Pr}[\sigma_A(0) = b]$ and $y_b = \mathbf{Pr}[\sigma_A(1) = b]$ for $b \in \{0, 1\}$. (So A's opting-out probabilities are $x_\perp = \mathbf{Pr}[\sigma_A(0) = \perp] = 1 - x_0 - x_1$ and $y_\perp = \mathbf{Pr}[\sigma_A(1) = \perp] = 1 - y_0 - y_1$.)

Recall, in addition to the conditions specifically stated in Case 6 in Table 1, 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, this condition implies that B's BNE strategy, represented by the point

$$P_5 = \left(\frac{D_0 D_1 M_{0,1} M_{1,0} - D_1^2 M_{0,1} M_{1,1}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}, \frac{D_0 D_1 M_{0,1} M_{1,0} - D_0^2 M_{0,0} M_{1,0}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}\right),$$

lies on the $p = q$ line. 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 his or her Nash strategy in order to make his or her opponent indifferent between various strategies rather than maximizing his or her own utility. Therefore, the coordinates of P_5 are such that they make agent A indifferent between opting out and playing $x_0 = 1$ (or opting out and $y_1 = 1$). 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 of P_5. (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 six cases in Table 1 holds.)

The rest of the section is devoted to proving that the six cases detailed in Table 1 indeed characterize all of the BNEs of the game under any possible setting of parameters. First, in Section 5.1 we argue that under the conditions of Theorem 5.1 agent B indeed has a BNE strategy which is

represented by the point P_5 in Table 1. In Section 5.2 we argue that in each of the cases detailed in Table 1 the detailed strategies form BNEs. Lastly, in Section 5.3 we argue completeness—that the six cases span all possible settings of parameters and that there are no additional BNEs.

5.1 Proof of Theorem 5.1: Finding a BNE Strategy for B

Recall, we assume $\mathbf{Pr}[t=0]=D_0$ and $\mathbf{Pr}[t=1]=D_1$, where wlog $D_0 \geq D_1$. As we did before, we denote B's strategy σ_B using $p=\mathbf{Pr}[\sigma_B(0)=0]$ and $q=\mathbf{Pr}[\sigma_B(1)=1]$. In contrast to the previous analysis, now A has to decide between three alternatives per $\hat{t}$ signal, so A has six options. However, seeing as A's choice to opt out always give A a utility of 0, we just denote four alternatives:

$$x_0=\mathbf{Pr}[\sigma_A(0)=0],\quad x_1=\mathbf{Pr}[\sigma_A(0)=1]$$
$$y_0=\mathbf{Pr}[\sigma_A(1)=0],\quad y_1=\mathbf{Pr}[\sigma_A(1)=1],$$

and we constrain $x_0+x_1\leq 1$ and $y_0+y_1\leq 1$.[10]

Now, given that A views a signal $\hat{t}$, he or she has three alternatives:

- Accuse B of being of type 0 and get an expected revenue of
$$\begin{aligned}&M_{0,0}\mathbf{Pr}[t=0\mid\hat{t}]-M_{0,1}\mathbf{Pr}[t=1\mid\hat{t}]\\&=\frac{1}{\mathbf{Pr}[\hat{t}]}\left(M_{0,0}\mathbf{Pr}[t=0]\mathbf{Pr}[\sigma_B(0)=\hat{t}]-M_{0,1}\mathbf{Pr}[t=1]\mathbf{Pr}[\sigma_B(1)=\hat{t}]\right)\\&=\frac{1}{\mathbf{Pr}[\hat{t}]}\left(M_{0,0}D_0\mathbf{Pr}[\sigma_B(0)=\hat{t}]-M_{0,1}D_1\mathbf{Pr}[\sigma_B(1)=\hat{t}]\right).\end{aligned}$$
- Accuse B of being of type 1 and get an expected revenue of
$$\begin{aligned}&M_{1,1}\mathbf{Pr}[t=1\mid\hat{t}]-M_{1,0}\mathbf{Pr}[t=0\mid\hat{t}]\\&=\frac{1}{\mathbf{Pr}[\hat{t}]}\left(M_{1,1}\mathbf{Pr}[t=1]\mathbf{Pr}[\sigma_B(1)=\hat{t}]-M_{1,0}\mathbf{Pr}[t=0]\mathbf{Pr}[\sigma_B(0)=\hat{t}]\right)\\&=\frac{1}{\mathbf{Pr}[\hat{t}]}\left(M_{1,1}D_1\mathbf{Pr}[\sigma_B(1)=\hat{t}]-M_{1,0}D_0\mathbf{Pr}[\sigma_B(0)=\hat{t}]\right).\end{aligned}$$
- Opt out and get revenue of $0=\frac{0}{\mathbf{Pr}[\hat{t}]}$.

This means that A prefers accusing B of being of type 0 to opting out when

$$\mathbf{Pr}[\sigma_B(0)=\hat{t}]>\frac{M_{0,1}D_1}{M_{0,0}D_0}\mathbf{Pr}[\sigma_B(1)=\hat{t}].$$

Similarly, A prefers accusing B of being of type 1 to opting out when

$$\mathbf{Pr}[\sigma_B(0)=\hat{t}]<\frac{M_{1,1}D_1}{M_{1,0}D_0}\mathbf{Pr}[\sigma_B(1)=\hat{t}].$$

From Equation (7) we have that $\frac{M_{1,1}D_1}{M_{1,0}D_0}<1<\frac{M_{0,1}D_1}{M_{0,0}D_0}$. Therefore, *given that* $\mathbf{Pr}[\hat{t}]>0$, A's best response is determined by the ratio:

$$\frac{\mathbf{Pr}[\sigma_B(0)=\hat{t}]}{\mathbf{Pr}[\sigma_B(1)=\hat{t}]}\begin{cases}<\frac{M_{1,1}D_1}{M_{1,0}D_0}, & A\text{ plays }\mathbf{Pr}[\sigma_A(\hat{t})=1]=1\\=\frac{M_{1,1}D_1}{M_{1,0}D_0}, & A\text{ is indifferent between }\perp\text{ and playing }\tilde{t}=1\\\in\left(\frac{M_{1,1}D_1}{M_{1,0}D_0},\frac{M_{0,1}D_1}{M_{0,0}D_0}\right) & A\text{ plays }\mathbf{Pr}[\sigma_A(\hat{t})=\perp]=1\\=\frac{M_{0,1}D_1}{M_{0,0}D_0}, & A\text{ is indifferent between }\perp\text{ and playing }\tilde{t}=0\\>\frac{M_{0,1}D_1}{M_{0,0}D_0} & A\text{ plays }\mathbf{Pr}[\sigma_A(\hat{t})=0]=1.\end{cases}$$

[10] Whereas in the previous section we constrained $x_0+x_1=1$ and $y_0+y_1=1$.

Therefore, A's BNE strategy when viewing the signal $\hat{t}$ (which is the best response to B's BNE strategy) is such that A never plays both $\tilde{t} = \hat{t}$ and $\tilde{t} = 1 - \hat{t}$ with nonzero probability.

Switching to the B agent, the utility functions of B are similar to before:

$$\text{For type } t = 0 : U_{B,0} = p(\rho_0 - x_0 M_{0,0} + x_1 M_{1,0}) + (1-p)(-y_0 M_{0,0} + y_1 M_{1,0}),$$
$$\text{For type } t = 1 : U_{B,1} = q(\rho_1 - y_1 M_{1,1} + y_0 M_{0,1}) + (1-q)(-x_1 M_{1,1} + x_0 M_{0,1}),$$

and so $p = 1$ if $\rho_0 > M_{0,0}(x_0 - y_0) - M_{1,0}(x_1 - y_1)$ and $p = 0$ if $\rho_0 < M_{0,0}(x_0 - y_0) - M_{1,0}(x_1 - y_1)$; similarly, $q = 1$ if $\rho_1 > M_{1,1}(y_1 - x_1) - M_{0,1}(y_0 - x_0)$ and $q = 0$ if $\rho_1 < M_{1,1}(y_1 - x_1) - M_{0,1}(y_0 - x_0)$.

We can now make our first claim about the BNE of the game.

Claim 5.2. *In any BNE strategy of B we have that either*

$$\frac{p}{1-q} = \frac{\mathbf{Pr}[\sigma_B^*(0) = 0]}{\mathbf{Pr}[\sigma_B^*(1) = 0]} \geq \frac{M_{0,1}D_1}{M_{0,0}D_0} \quad or \quad \frac{1-p}{q} = \frac{\mathbf{Pr}[\sigma_B^*(0) = 1]}{\mathbf{Pr}[\sigma_B^*(1) = 1]} \leq \frac{M_{1,1}D_1}{M_{1,0}D_0}.$$

Proof. Assume for the sake of contradiction that both conditions do not hold. Then, given the $\hat{t} = 0$ signal, it holds that $x_0 = \mathbf{Pr}[\sigma_A^*(0) = 0] = 0$, and given the $\hat{t} = 1$ signal, it holds that $y_1 = \mathbf{Pr}[\sigma_A^*(1) = 1] = 0$. Thus, B's best response to A's strategy is to switch to $(p, q) = (1, 1)$ (since $\rho_0, \rho_1 > 0$) and now both conditions do hold. □

Claim 5.3. *In any BNE strategy of B we have that both*

$$\frac{p}{1-q} = \frac{\mathbf{Pr}[\sigma_B^*(0) = 0]}{\mathbf{Pr}[\sigma_B^*(1) = 0]} > \frac{M_{1,1}D_1}{M_{1,0}D_0} \quad and \quad \frac{1-p}{q} = \frac{\mathbf{Pr}[\sigma_B^*(0) = 1]}{\mathbf{Pr}[\sigma_B^*(1) = 1]} < \frac{M_{0,1}D_1}{M_{0,0}D_0}.$$

Proof. Based on the previous claim, one of the two inequalities is immediate. Assume we have $\frac{p}{1-q} \geq \frac{M_{0,1}D_1}{M_{0,0}D_0} > 1 > \frac{M_{1,1}D_1}{M_{1,0}D_0}$; we now show that $\frac{1-p}{q} < \frac{M_{0,1}D_1}{M_{0,0}D_0}$ must also hold. If, for contradiction, we have that $\frac{1-p}{q} \geq \frac{M_{0,1}D_1}{M_{0,0}D_0}$, then

$$1 = p + (1-p) \geq \frac{M_{0,1}D_1}{M_{0,0}D_0}(q + (1-q)) = \frac{M_{0,1}D_1}{M_{0,0}D_0},$$

which contradicts Equation (7). The argument for the case $\frac{1-p}{q} \leq \frac{M_{1,1}D_1}{M_{1,0}D_0}$ is symmetric. □

Based on the last claim and on A's best-response analysis, we have that in any BNE strategy of A it holds that $x_1 = \mathbf{Pr}[\sigma_A^*(0) = 1] = 0$ and $y_0 = \mathbf{Pr}[\sigma_A^*(1) = 0] = 0$ (i.e., given the signal $\hat{t}$, A never plays $\tilde{t} = 1 - \hat{t}$). As a result, B's best-response analysis simplifies to $p = 1$ if $\rho_0 > M_{0,0}x_0 + M_{1,0}y_1$ and $p = 0$ if $\rho_0 < M_{0,0}x_0 + M_{1,0}y_1$; similarly, $q = 1$ if $\rho_1 > M_{1,1}y_1 + M_{0,1}x_0$ and $q = 0$ if $\rho_1 < M_{1,1}y_1 + M_{0,1}x_0$.

We are now able to prove the existence of a BNE as specified in Theorem 5.1.

Claim 5.4. *Assume that $0 \leq \rho_1 M_{1,0} - \rho_0 M_{1,1} \leq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$ and $0 \leq \rho_0 M_{0,1} - \rho_1 M_{0,0} \leq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$. The strategies σ_A^* and σ_B^* denoted below are BNE strategies.*

$$\begin{aligned}
\text{For } A: x_0^* &= \mathbf{Pr}[\sigma_A^*(0) = 0] = \frac{M_{1,0}\rho_1 - M_{1,1}\rho_0}{M_{1,0}M_{0,1} - M_{0,0}M_{1,1}} \\
x_1^* &= \mathbf{Pr}[\sigma_A^*(0) = 1] = 0 \\
y_0^* &= \mathbf{Pr}[\sigma_A^*(1) = 0] = 0 \\
y_1^* &= \mathbf{Pr}[\sigma_A^*(1) = 1] = \frac{M_{0,1}\rho_0 - M_{0,0}\rho_1}{M_{1,0}M_{0,1} - M_{0,0}M_{1,1}}
\end{aligned}$$

$$
\begin{aligned}
\text{For } B: p^* &= \mathbf{Pr}[\sigma_B^*(0) = 0] = \frac{D_1 M_{0,1}(D_0 M_{1,0} - D_1 M_{1,1})}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}} \\
1 - p^* &= \mathbf{Pr}[\sigma_B^*(0) = 1] = \frac{D_1 M_{1,1}(D_1 M_{0,1} - D_0 M_{0,0})}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}} \\
1 - q^* &= \mathbf{Pr}[\sigma_B^*(1) = 0] = \frac{D_0 M_{0,0}(D_0 M_{1,0} - D_1 M_{1,1})}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}} \\
q^* &= \mathbf{Pr}[\sigma_B^*(1) = 1] = \frac{D_0 M_{1,0}(D_1 M_{0,1} - D_0 M_{0,0})}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}
\end{aligned}
$$

Proof. First, observe that under the given assumptions in the claim it holds that $x_0^*, y_1^* \in [0,1]$, and due to Equation (7) it holds that $p^*, q^*, 1-p^*, 1-q^*$ are all strictly positive (so $p^*, q^* \in (0,1)$).

Now observe that when B follows σ_B^*, A has no incentive to deviate since $\frac{p^*}{1-q^*} = \frac{M_{0,1}D_1}{M_{0,0}D_0}$ and $\frac{1-p^*}{q^*} = \frac{M_{1,1}D_1}{M_{1,0}D_0}$. When A follows σ_A^*, B has no incentive to deviate since

$$
\begin{aligned}
M_{0,0}x_0^* + M_{1,0}y_1^* &= \frac{M_{1,0}M_{0,1}\rho_0 - M_{0,0}M_{1,1}\rho_0}{M_{1,0}M_{0,1} - M_{0,0}M_{1,1}} &&= \rho_0 \\
M_{0,1}x_0^* + M_{1,1}y_1^* &= \frac{M_{0,1}M_{1,0}\rho_1 - M_{1,1}M_{0,0}\rho_1}{M_{1,0}M_{0,1} - M_{0,0}M_{1,1}} &&= \rho_1.
\end{aligned}
$$

□

Observe that when $D_0^2 M_{0,0} M_{1,0} = D_1^2 M_{0,1} M_{1,1}$, $p^* = q^*$. Furthermore, in this case we have that $p^* > \frac{1}{2}$ because

$$
\begin{aligned}
&2D_0 D_1 M_{0,1} M_{1,0} - 2D_1^2 M_{0,1} M_{1,1} > D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1} \\
&\Leftrightarrow D_0 D_1 M_{0,1} M_{1,0} - D_1^2 M_{0,1} M_{1,1} > D_0^2 M_{0,0} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1} \\
&\Leftrightarrow D_1 M_{0,1}(D_0 M_{1,0} - D_1 M_{1,1}) > D_0 M_{0,0}(D_0 M_{1,0} - D_1 M_{1,1}) \\
&\Leftrightarrow 0 > -D_1 M_{0,1} + D_0 M_{0,0},
\end{aligned}
$$

where the last derivation and the last inequality are both true because of Equation (7). This concludes the existence part of Theorem 5.1. The more complicated part is to show that B's BNE strategy is *unique*. Formally, our goal is to prove the following.

Theorem 5.5. *Assume that* $0 < \rho_1 M_{1,0} - \rho_0 M_{1,1} < M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$ *and* $0 < \rho_0 M_{0,1} - \rho_1 M_{0,0} < M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$. *Then in all BNEs of the game both types of* B *agents play a mixed strategy.*

Assuming Theorem 5.5 holds and using our above best-response analysis, the uniqueness of the BNE of Theorem 5.1 is immediate. If both p and q are nonintegral, then it must hold that x_0^* and y_1^* are unique, since this pair is the unique solution to a well-defined system of two linear equations in two variables that set both types of B agents indifferent. Under the assumption of Theorem 5.5, we have that both x_0^* and y_1^* are nonintegral as well. This means that B plays (p,q) s.t. A is indifferent to the value of x_0^*, y_1^*, i.e., $p = (1-q)\frac{M_{0,1}D_1}{M_{0,0}D_0}$ and $1-p = q\frac{M_{1,1}D_1}{M_{1,0}D_0}$. Again, since this is a linear system in two variables, there exists a unique (p,q) pair that satisfies this condition, which is given by (p^*, q^*).

In the remainder of this section, our goal is to prove Theorem 5.5. In fact, we give a full analysis of all the points (p,q) that *may* be B's BNE strategy, and for each such possible (p,q) we analyze the conditions over the parameters of the game under which it is a BNE strategy for B. The analysis is fairly long and tedious, as it involves checking feasibility constraints over the six parameters

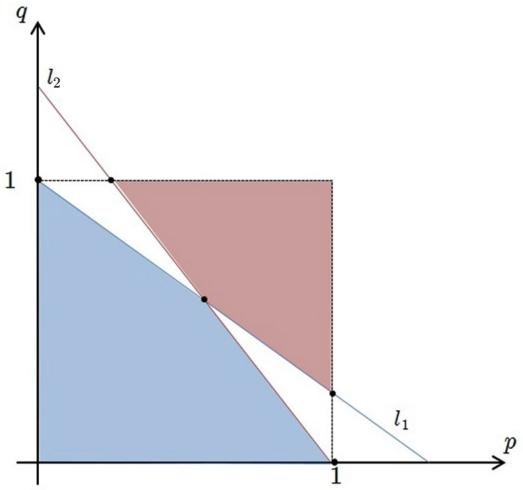

Fig. 3. The four regions created by the l_1 and l_2, and the five intersection points of the two lines and the borders of the squares.

of the game: ρ_0, ρ_1, $M_{0,0}$, $M_{0,1}$, $M_{1,0}$, and $M_{1,1}$. Furthermore, after deriving the suitable feasibility constraints, we show (in Section 5.3) that they cover all settings of the parameters of the game and are mutually exclusive (when inequalities are strict).

5.2 Proof of Theorem 5.1: Characterizing All Potential BNEs of the Game

Consider the space $[0,1] \times [0,1]$ of all possible strategies (p,q) for the two types of B agents. We denote the two "lines of indifference" for A on this square:

$$l_1 : p = \frac{M_{0,1}D_1}{M_{0,0}D_0}(1-q)$$
$$l_2 : \ 1-p = \frac{M_{1,1}D_1}{M_{0,1}D_1}q,$$

where $(p,q) = (0,1) \in l_1$ and $(p,q) = (1,0) \in l_2$. These lines partition the $[0,1] \times [0,1]$ square into multiple different regions, as shown in Figure 3.

First, we argue that any (p,q) in the lower region, below l_1 and l_2 (the shaded blue region in Figure 3), cannot be B's strategy in a BNE. We in fact have already shown this: for any (p,q) in this blue region we have that $\frac{p}{1-q} < \frac{M_{0,1}D_1}{M_{0,0}D_0}$ and $\frac{1-p}{q} < \frac{M_{1,1}D_1}{M_{0,1}D_1}$, contradicting our earlier claim.

Second, we argue that a point (p,q) above both lines is B's strategy in a BNE if the valuation of B for the coupons is high. Observe that, for any (p,q) above the line, A's best response is to set $x_0 = y_1 = 1$, which means B's utility is $p(\rho_0 - M_{0,0}) + (1-p)M_{1,0}$ for agents of type $t = 0$, and $q(\rho_1 - M_{1,1}) + (1-q)M_{0,1}$ for type $t = 1$. Therefore, B has no incentive to deviate from (p,q) only if $\rho_0 \geq M_{0,0} + M_{1,0}$ and $\rho_1 \geq M_{0,1} + M_{1,1}$. In particular, when both inequalities are strict, we have that $(1,1)$ is B's BNE; when both are equalities, any point above both l_1 and l_2 is a BNE; and when one is an equality and the other is a strict inequality, we have that the BNE strategy is on the border of the $[0,1] \times [0,1]$ square.

We now turn our attention to the points strictly between the lines l_1 and l_2 (excluding all points on these lines). To any (p,q) in these regions, A's best response is to play $\tilde{t} = \hat{t}$ when seeing one signal and to opt out when seeing the other signal; e.g., for any (p,q) above l_1 but below l_2 (the top-left region), A opts out when seeing the $\hat{t} = 1$ signal but plays $\tilde{t} = 0$ when seeing the $\hat{t} = 0$ signal. In that case, B's utility function is $p(\rho_0 - M_{0,0})$ for type $t = 0$, and $q\rho_1 + (1-q)M_{0,1}$ for type $t = 1$.

Therefore, unless $\rho_0 = M_{0,0}$, agents of type $t = 0$ have incentive to deviate (either to playing $p = 0$ or $p = 1$). In addition, it must also hold that $\rho_1 \geq M_{0,1}$. (If this inequality is strict, then the BNE strategy lies on the border of the square.) Analogously, should the BNE strategy lie above the l_2 line but below the l_1 line (lower-right area), then it must be the case that $\rho_1 = M_{1,1}$ and $\rho_0 \geq M_{1,0}$.

We now consider points on the l_1 and l_2, excluding the five intersection points we have with either of the two lines or any of the lines intersecting with the $p = 1$ line or the $q = 1$ line.

(i) For any (p, q) on the l_1 line below the l_2 line (top-left side of l_1 bordering the blue region): A's best response to any such (p, q) is to set $y_0 = y_1 = 0$ and $x_1 = 0$. It follows that B's utility is $p(\rho_0 - x_0 M_{0,0})$ and $q\rho_1 + (1-q)x_0 M_{0,1}$ for types $t = 0$ and $t = 1$, respectively. Hence, if $0 \leq \frac{\rho_0}{M_{0,0}} = \frac{\rho_1}{M_{0,1}} \leq 1$, then such strategies can be BNE.

(ii) For any (p, q) on the l_1 line above the l_2 line (bottom-right side of l_1 bordering the red region): A's best response to such (p, q) is to set $x_1 = y_0 = 0$, $y_1 = 1$; B's utility functions are $p(\rho_0 - x_0 M_{0,0}) + (1-p)M_{1,0}$ and $q(\rho_1 - M_{1,1}) + (1-q)x_0 M_{0,1}$ for types $t = 0$ and $t = 1$, respectively. It follows that if $0 \leq \frac{\rho_0 - M_{1,0}}{M_{0,0}} = \frac{\rho_1 - M_{1,1}}{M_{0,1}} \leq 1$, then such point gives a BNE.

(iii) For any (p, q) on the l_2 line below the l_1 line (bottom-right side of l_2 bordering the blue region): As a response to any (p, q) here, A sets $x_0 = x_1 = 0$ and $y_0 = 0$. B's utility function is therefore $p\rho_0 + (1-p)(y_1 M_{1,0})$ and $q(\rho_1 - y_1 M_{1,1})$. Hence, in case $0 \leq \frac{\rho_0}{M_{1,0}} = \frac{\rho_1}{M_{1,1}} \leq 1$, we have a BNE with such (p, q).

(iv) For any (p, q) on the l_2 line above the l_1 line (top-left side of l_2 bordering the red region): As best response to such (p, q), A sets $x_1 = y_0 = 1$ and $x_0 = 1$. And so B's utility functions are $p(\rho_0 - M_{0,0}) + (1-p)y_1 M_{1,0}$ and $q(\rho_1 - y_1 M_{1,1} + (1-q)M_{0,1}$. Therefore, if we have $0 \leq \frac{\rho_0 - M_{0,0}}{M_{1,0}} = \frac{\rho_1 - M_{0,1}}{M_{1,1}} \leq 1$, then we have a BNE with such (p, q).

Thus far (with the exception of the potential BNE at the point $(p, q) = (1, 1)$) we have considered only BNEs that may arise only when the parameters of the game (ρ_0, ρ_1 and the entries of M) satisfy some equality constraints. Assuming we perturb the values of ρ_0 and ρ_1 a little such that none of the above-mentioned equalities hold, we are left with five points on which the BNE can occur:

$$P_1 = (0, 1), P_2 = \left(1 - \frac{D_1 M_{1,1}}{D_0 M_{1,0}}, 1\right), P_3 = (1, 0), P_4 = \left(1, 1 - \frac{D_0 M_{0,0}}{D_1 M_{0,1}}\right)$$

$$P_5 = \left(\frac{D_0 D_1 M_{0,1} M_{1,0} - D_1^2 M_{0,1} M_{1,1}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}, \frac{D_0 D_1 M_{0,1} M_{1,0} - D_0^2 M_{0,0} M_{1,0}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}\right).$$

We traverse them one by one. We remind the reader that since $\frac{M_{0,1}}{M_{0,0}} > \frac{D_0}{D_1} > \frac{M_{1,1}}{M_{1,0}}$, $M_{0,0} M_{1,1} < M_{0,1} M_{1,0}$. We repeatedly use this inequality in the analysis below.

5.2.1 Conditions under Which B's BNE Strategy Is P_1. Observe that in this case B never sends the $\hat{t} = 0$ signal. A's best response is naturally to opt out, but A still commits to certain values of x_0 and x_1 (to prevent B from deviating from the $(0, 1)$ strategy). B's utility functions are

$$\text{For } t = 0: \ p(\rho_0 - x_0 M_{0,0} + x_1 M_{1,0}).$$
$$\text{For } t = 1: \ q\rho_1 + (1-q)(x_0 M_{0,1} - x_1 M_{1,1}).$$

So A should set x_0 and x_1 s.t. $\rho_0 \leq x_0 M_{0,0} - x_1 M_{1,0}$ and $\rho_1 \geq x_0 M_{0,1} - x_1 M_{1,1}$.

Proposition 5.6. *There exist $x_0, x_1 \in [0, 1]$ satisfying both $\rho_0 \leq x_0 M_{0,0} - x_1 M_{1,0}$ and $\rho_1 \geq x_0 M_{0,1} - x_1 M_{1,1}$ iff $\rho_0 \leq M_{0,0}$ and $\rho_0 M_{0,1} \leq \rho_1 M_{0,0}$.*

Proof. To see that these conditions are sufficient, assume that $\rho_0 \le M_{0,0}$ and $\rho_0 M_{0,1} \le \rho_1 M_{0,0}$. Then we can set $x_0 = \frac{\rho_0}{M_{0,0}}$ and $x_1 = 0$. Clearly, both lie on the $[0,1]$-interval. We can check and see that indeed $\rho_0 \le \frac{\rho_0}{M_{0,0}} M_{0,0} - 0 \cdot M_{1,0}$ and $\rho_1 \ge \frac{\rho_0}{M_{0,0}} M_{0,1} - 0 \cdot M_{1,1}$.

We now show these conditions are necessary. Suppose that $\rho_0 > M_{0,0}$; then observe that any x_0, x_1 satisfying the two constraints must satisfy $0 \le x_1 M_{1,0} \le x_0 M_{0,0} - \rho_0$, so $x_0 \ge \frac{\rho_0}{M_{0,0}}$. As a result of our assumption, we have that $x_0 > 1$. Contradiction.

So assume now that $\rho_0 \le M_{0,0}$ yet $\rho_0 M_{0,1} > \rho_1 M_{0,0}$. Any x_0, x_1 satisfying the two constraints must also satisfy

$$x_0 \frac{M_{0,1}}{M_{1,1}} - \frac{\rho_1}{M_{1,1}} \le x_1 \le x_0 \frac{M_{0,0}}{M_{1,0}} - \frac{\rho_0}{M_{1,0}},$$

which, using our assumption, yields

$$x_0 \frac{M_{0,1}M_{1,0} - M_{0,0}M_{1,1}}{M_{1,1}M_{1,0}} \le \frac{\rho_1}{M_{1,1}} - \frac{\rho_0}{M_{1,0}} < \rho_0 \left(\frac{M_{0,1}}{M_{0,0}M_{1,1}} - \frac{1}{M_{1,0}} \right) = \rho_0 \frac{M_{1,0}M_{0,1} - M_{0,0}M_{1,1}}{M_{0,0}M_{1,0}M_{1,1}},$$

so we have $x_0 < \frac{\rho_0}{M_{0,0}}$. Contradiction. □

5.2.2 Conditions under Which B's BNE Strategy Is P_2. As a response to this strategy, A's best response is to set $x_1 = y_0 = 0$ and $x_0 = 1$, while y_1 is to be determined. So B's utility functions are

$$\text{For } t = 0: \ p(\rho_0 - M_{0,0}) + (1-p) y_1 M_{1,0}.$$
$$\text{For } t = 1: \ q(\rho_1 - y_1 M_{1,1}) + (1-q) M_{0,1}.$$

Therefore, in order for B to not have any incentive to deviate, A should set y_1 s.t $\rho_0 - M_{0,0} = y_1 M_{1,0}$ and $\rho_1 - y_1 M_{1,1} \ge M_{0,1}$.

Proposition 5.7. *There exists a $y_1 \in [0,1]$ satisfying both $\rho_0 - M_{0,0} = y_1 M_{1,0}$ and $\rho_1 - y_1 M_{1,1} \ge M_{0,1}$ iff $M_{0,0} \le \rho_0 \le M_{0,0} + M_{1,0}$ and $(\rho_0 - M_{0,0}) M_{1,1} \le (\rho_1 - M_{0,1}) M_{1,0}$.*

Proof. Clearly, the only y_1 that can satisfy both constraints is $y_1 = \frac{\rho_0 - M_{0,0}}{M_{1,0}}$, and we therefore must have that $M_{0,0} \le \rho_0 \le M_{0,0} + M_{1,0}$. We also need to verify that indeed the inequality holds in the right direction, i.e., to have $(\rho_0 - M_{0,0}) \frac{M_{1,1}}{M_{1,0}} \le \rho_1 - M_{0,1}$. Clearly, if those two conditions hold, then y_1 defined as above satisfies the required. □

Observation 5.8. *If we have that $(\rho_0 - M_{0,0}) M_{1,1} \le (\rho_1 - M_{0,1}) M_{1,0}$, then also $\frac{\rho_0}{\rho_1} < \frac{M_{1,0}}{M_{1,1}}$.*

Proof. $\rho_0 M_{1,1} - M_{0,0} M_{1,1} \le \rho_1 M_{1,0} - M_{0,1} M_{1,0} < \rho_1 M_{1,0} - M_{0,0} M_{1,1} \Rightarrow \rho_0 M_{1,1} < \rho_1 M_{1,0}$. □

5.2.3 Conditions ynder Which B's BNE Strategy Is P_3. Should B play $(1,0)$, then we have that A only sees the $\hat{t} = 0$ signal and always opts out (i.e., $x_0 = x_1 = 0$). However, in order to prevent B from deviating, A needs to commit to a y_0, y_1 that leaves B preferring not to deviate from $(1,0)$. B's utility functions are

$$\text{For } t = 0: \ p\rho_0 + (1-p)(-y_0 M_{0,0} + y_1 M_{1,0}).$$
$$\text{For } t = 1: \ q(\rho_1 + y_0 M_{0,1} - y_1 M_{1,1}).$$

Therefore, in order for B to not have any incentive to deviate, A should set y_0, y_1 s.t $\rho_0 \ge -y_0 M_{0,0} + y_1 M_{1,0}$ and $\rho_1 + y_0 M_{0,1} - y_1 M_{1,1} \le 0$.

Proposition 5.9. *There exist $y_0, y_1 \in [0,1]$ satisfying both $\rho_0 \ge -y_0 M_{0,0} + y_1 M_{1,0}$ and $\rho_1 \le -y_0 M_{0,1} + y_1 M_{1,1}$ iff $\rho_1 \le M_{1,1}$ and $\rho_0 M_{1,1} \ge \rho_1 M_{1,0}$.*

The proof is completely analogous to the proof of Proposition 5.6.

5.2.4 Conditions under Which B's BNE Strategy Is P_4. As a response to this strategy, A's best response is to set $x_1 = y_0 = 0$ and $y_1 = 1$, while x_0 is to be determined. So B's utility functions are

$$\text{For } t = 0: \ p(\rho_0 - x_0 M_{0,0}) + (1-p)M_{1,0}.$$
$$\text{For } t = 1: \ q(\rho_1 - M_{1,1}) + (1-q)x_0 M_{0,1}.$$

Therefore, in order for B to not have any incentive to deviate, A should set x_0 s.t $\rho_0 - x_0 M_{0,0} \geq M_{1,0}$ and $\rho_1 - M_{1,1} = x_0 M_{0,1}$.

Proposition 5.10. *There exists a $x_0 \in [0,1]$ satisfying both $\rho_0 - x_0 M_{0,0} \geq M_{1,0}$ and $\rho_1 - M_{1,1} = x_0 M_{0,1}$ iff $M_{1,1} \leq \rho_1 \leq M_{1,1} + M_{0,1}$ and $(\rho_1 - M_{1,1})\, M_{0,0} \leq (\rho_0 - M_{1,0})\, M_{0,1}$.*

The proof is analogous to the proof of Proposition 5.7.

5.2.5 Conditions under Which B's BNE Strategy Is P_5. As this point lies on the intersection of l_1 and l_2, A's best response to this strategy is to set $x_1 = y_0 = 0$. Thus, B's utility functions are

$$\text{For } t = 0: \ p(\rho_0 - x_0 M_{0,0}) + (1-p)y_1 M_{1,0}.$$
$$\text{For } t = 1: \ q(\rho_1 - y_1 M_{1,1}) + (1-q)x_0 M_{0,1}.$$

It is therefore up to A to pick x_0 and y_1 that satisfy both equalities $\binom{\rho_0}{\rho_1} = \begin{pmatrix} M_{0,0} & M_{1,0} \\ M_{0,1} & M_{1,1} \end{pmatrix} \binom{x_0}{y_1}$. Cramer's formula give that the solution to this system is

$$\begin{pmatrix} x_0 \\ y_1 \end{pmatrix} = \frac{1}{M_{1,0}M_{0,1} - M_{0,0}M_{1,1}} \begin{pmatrix} -M_{1,1} & M_{1,0} \\ M_{0,1} & -M_{0,0} \end{pmatrix} \begin{pmatrix} \rho_0 \\ \rho_1 \end{pmatrix}. \tag{9}$$

In order for x_0, y_1 to be in the range $[0,1]$, we therefore must have that (1) $\rho_0 M_{1,1} \leq \rho_1 M_{1,0}$, (2) $(\rho_1 - M_{0,1})M_{1,0} \leq (\rho_0 - M_{0,0})M_{1,1}$, (3) $\rho_1 M_{0,0} \leq \rho_0 M_{0,1}$, and (4) $(\rho_0 - M_{1,0})M_{0,1} \leq (\rho_1 - M_{1,1})M_{0,0}$. In other words:

$$0 \leq \rho_1 M_{1,0} - \rho_0 M_{1,1} \leq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$$
$$0 \leq \rho_0 M_{0,1} - \rho_1 M_{0,0} \leq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}. \tag{10}$$

5.2.6 Summarizing. Table 1 summarized the six conditions under which each point is a BNE. It is easy to verify that the condition of Theorem 5.5 is precisely condition 6, underwhich B's BNE strategy is unique (the point P_5) and it is mixed.

5.3 Proof of Theorem 5.1: The Uniqueness of B's BNE Strategy

So far we have introduced conditions for the existence of various BNEs. In this section, our goal is to show that the above analysis gives a complete description of the game, that is, to show that the cases detailed in Table 1 span all potential values the parameters of the game may take, and furthermore (modulo cases of equality between parameters) they are also mutually exclusive.

Lemma 5.11. *Assume that the parameters of the game (i.e., ρ_0, ρ_1 and the entries of M) satisfy one of the six conditions detailed in Table 1 with strict inequalities. Then no other condition in Table 1 holds simultaneously. In other words, the conditions in Table 1 are mutually exclusive (excluding equalities). As the conditions are mutually exclusive, it means that under the condition specified in Theorem 5.5, the game has a unique BNE—as specified by case 6 in Table 1.*

Proof. We traverse the six cases, showing that if case i holds with strict inequalities, then some other case $j > i$ cannot hold.

Case 1. Clearly, if the conditions of case 1 hold, then the conditions of cases 2, 3, 4, and 5 cannot hold. To see that the conditions of case 6 cannot hold, we argue that the condition $\max\{\rho_1 M_{1,0} - \rho_0 M_{1,1}, \rho_0 M_{0,1} - \rho_1 M_{0,0}\} \leq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$ implies that both $\rho_0 \leq M_{0,0} + M_{1,0}$ and $\rho_1 \leq M_{0,1} + M_{1,1}$. This claim follows from the inequalities

$$\begin{aligned}\rho_0(M_{0,1}M_{1,0} - M_{0,0}M_{1,1}) &= M_{0,0}(\rho_1 M_{1,0} - \rho_0 M_{1,1}) + M_{1,0}(\rho_0 M_{0,1} - \rho_1 M_{0,0}) \\ &\leq (M_{0,0} + M_{1,0})(M_{0,1}M_{1,0} - M_{0,0}M_{1,1}) \\ \rho_1(M_{0,1}M_{1,0} - M_{0,0}M_{1,1}) &= M_{0,1}(\rho_1 M_{1,0} - \rho_0 M_{1,1}) + M_{1,1}(\rho_0 M_{0,1} - \rho_1 M_{0,0}) \\ &\leq (M_{0,1} + M_{1,1})(M_{0,1}M_{1,0} - M_{0,0}M_{1,1}).\end{aligned}$$

Case 2. Clearly, the conditions of case 2 cannot hold simultaneously with the conditions of cases 4 and 6. To exclude the other cases, observe that using our favorite inequality $\frac{M_{0,0}}{M_{0,1}} < \frac{M_{1,0}}{M_{1,1}}$, we have that the condition $\frac{\rho_0}{\rho_1} \leq \frac{M_{0,0}}{M_{0,1}}$ implies that $\frac{\rho_0}{\rho_1} < \frac{M_{1,0}}{M_{1,1}}$. Hence, case 3 cannot hold, and neither does case 5 (using again the fact that $M_{0,1}M_{1,0} > M_{0,0}M_{1,1}$).

Case 3. This case is symmetric to case 2—since $M_{0,1}M_{1,0} - M_{0,0}M_{1,1} > 0$, case 3 rules out case 4 (and the fact that it cannot hold simultaneously with cases 5 and 6 is obvious).

Case 4. Clearly, case 6 cannot hold together with case 4. To show that case 5 cannot hold too, we claim that if both $\rho_1 M_{1,0} - \rho_0 M_{1,1} \geq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$ and $\rho_0 M_{0,1} - \rho_1 M_{0,0} \geq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$ hold, then $\rho_0 \geq M_{0,0} + M_{1,0}$. This holds because the two inequalities imply

$$\begin{aligned}&\rho_1 \leq (\rho_0 - M_{1,0})\frac{M_{0,1}}{M_{0,0}} + M_{1,1} \text{ and } \rho_1 \geq (\rho_0 - M_{0,0})\frac{M_{1,1}}{M_{1,0}} + M_{0,1} \\ \Rightarrow\ &\rho_0\left(\frac{M_{0,1}}{M_{0,0}} - \frac{M_{1,1}}{M_{1,0}}\right) \geq M_{0,1} - M_{1,1} + \frac{M_{1,0}M_{0,1}}{M_{0,0}} - \frac{M_{0,0}M_{1,1}}{M_{1,0}} \\ \Rightarrow\ &\rho_0 \geq \frac{M_{0,0}M_{0,1}M_{0,1} - M_{0,0}M_{1,0}M_{1,1} + M_{0,1}M_{1,0}^2 - M_{0,0}^2 M_{1,1}}{M_{0,1}M_{1,0} - M_{0,0}M_{1,1}} = M_{0,0} + M_{1,0}.\end{aligned}$$

Case 5. Clearly, cases 5 and 6 cannot hold simultaneously. □

Lemma 5.12. *Any choice of parameters for ρ_0, ρ_1 and the entries of M satisfies at least one of the six cases detailed in Table 1.*

Proof. First, suppose $\rho_0 \geq M_{0,0} + M_{1,0}$. We claim that in this case, the value of ρ_1 determines which case holds.

- If $\rho_1 \leq M_{1,1}$, then case 3 holds, since obviously $M_{1,1}\frac{\rho_0}{M_{1,0}} > M_{1,1} \geq \rho_1$.
- If $M_{1,1} < \rho_1 \leq M_{0,1} + M_{1,1}$, then case 5 holds since

$$\begin{aligned}\rho_0 M_{0,1} - \rho_1 M_{0,0} &\geq (M_{0,0} + M_{1,0})M_{0,1} - \rho_1 M_{0,0} \\ &= M_{0,1}M_{1,0} + M_{0,0}(M_{0,1} - \rho_1) \\ &\geq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}.\end{aligned}$$

- If $\rho_1 > M_{0,1} + M_{1,1}$, then clearly case 1 holds.

Symmetrically, if we have that $\rho_1 \geq M_{0,1} + M_{1,1}$, then the value of ρ_0 determines whether we fall into case 2, case 4, or case 1.

So assume from now on that both $\rho_0 < M_{0,0} + M_{1,0}$ and $\rho_1 < M_{0,1} + M_{1,1}$.

Suppose that in addition to these two upper bounds on the value of the coupon, we also have that $\frac{\rho_0}{\rho_1} \leq \frac{M_{0,0}}{M_{0,1}}$.

- If $\rho_0 \leq M_{0,0}$, then clearly case 2 holds.
- If $\rho_0 \geq M_{0,0}$, then we show that case 4 holds. Observe $\frac{\rho_1}{\rho_0} - \frac{M_{1,1}}{M_{1,0}} \geq \frac{M_{0,1}}{M_{0,0}} - \frac{M_{1,1}}{M_{1,0}}$, so $\frac{\rho_1 M_{1,0} - \rho_0 M_{1,1}}{\rho_0 M_{1,0}} \geq \frac{M_{0,1}M_{1,0} - M_{0,0}M_{1,1}}{M_{0,0}M_{1,0}}$. We conclude that $\rho_1 M_{1,0} - \rho_0 M_{1,1} \geq \frac{\rho_0}{M_{0,0}}(M_{0,1}M_{1,0} - M_{0,0}M_{1,1})$. So the fact that $\rho_0 \geq M_{0,0}$ implies that the conditions of case 4 hold.

And symmetrically, if we assume that $\rho_0 < M_{0,0} + M_{1,0}$ and $\rho_1 < M_{0,1} + M_{1,1}$ and that in addition $\frac{\rho_0}{\rho_1} \geq \frac{M_{1,0}}{M_{1,1}}$, then the same line of argument shows that either case 3 or case 5 holds.

So now, we assume both that $\rho_0 < M_{0,0} + M_{1,0}$, $\rho_1 < M_{0,1} + M_{1,1}$ and that $\frac{M_{0,0}}{M_{0,1}} < \frac{\rho_0}{\rho_1} < \frac{M_{1,0}}{M_{1,1}}$.

- If $\rho_1 M_{1,0} - \rho_0 M_{1,1} \geq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$, we argue that case 3 holds. This is because we have both that $\rho_1 < \rho_0 \frac{M_{0,1}}{M_{0,0}}$ and that $\rho_1 \geq \rho_0 \frac{M_{1,1}}{M_{1,0}} + M_{0,1} - \frac{M_{0,0}M_{1,1}}{M_{1,0}}$. Combining the two, we get
$$\rho_0 \left(\frac{M_{0,1}}{M_{0,0}} - \frac{M_{1,1}}{M_{1,0}}\right) > \frac{M_{0,1}M_{1,0} - M_{0,0}M_{1,1}}{M_{1,0}} \Rightarrow \rho_0 > M_{0,0}.$$
- If $\rho_0 M_{0,1} - \rho_1 M_{0,0} \geq M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$, then we are in the analogous case (case 5), as we can show, using the inequality $\frac{\rho_0}{\rho_1} < \frac{M_{1,0}}{M_{1,1}}$, that $\rho_1 > M_{1,1}$.

This leaves us with the case that $\rho_0 < M_{0,0} + M_{1,0}$, $\rho_1 < M_{0,1} + M_{1,1}$, $\frac{M_{0,0}}{M_{0,1}} < \frac{\rho_0}{\rho_1} < \frac{M_{1,0}}{M_{1,1}}$ and also $\rho_1 M_{1,0} - \rho_0 M_{1,1} < M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$ and $\rho_0 M_{0,1} - \rho_1 M_{0,0} < M_{0,1}M_{1,0} - M_{0,0}M_{1,1}$. This is precisely case 6. □

6 CONCLUSIONS AND FUTURE DIRECTIONS

Our work is a first attempt at exposing and reconciling the competing conclusions of two different approaches to the same challenge: the theory of privacy-aware agents (where privacy loss is modeled using differential privacy) and the behavior of standard utility-maximizing agents once they explicitly assess future losses from having their behavior in the current game publicly exposed. While the canonical privacy-aware agent randomizes his or her strategy, we show that different explicit privacy losses cause very different behavior among agents. This is best illustrated with the game studied in Section 4 (Theorem 4.2). In that game, agents assess their future loss and their behavior is therefore quite simple: if the current gain is greater than the future loss, their behavior is to truthfully report their type; otherwise, they lie and report the opposite type. We believe this simple rule explains real-life phenomena, such as people trying to hide their medical condition from the general public while truthfully answering a doctor's questions.[11]

Observe, however, that in all the games we analyzed, we still have not pinned down a game in which the behavior of a non-privacy-aware agent *fully* mimics the behavior of a privacy-aware agent. Privacy-aware agents' behavior is, after a fashion, quite reasonable. They trade off between the value of the coupon they get and the amount of privacy (or change in belief) they are willing to risk. Naturally, the higher the value of the coupon, the more privacy they are willing to risk. In contrast, in the game discussed in Section 5, even under settings where B's BNE strategy σ_B^* is randomized and satisfies $\mathbf{Pr}[\sigma_B^*(0)] = \mathbf{Pr}[\sigma_B^*(1)]$, we don't see a continuous change in B's behavior based on the value of the coupon. Changing solely the value of the coupon while keeping all other parameters the same, we see that B plays the same BNE strategy, whereas A's BNE strategy continuously changes.

A natural follow-up question is to extend our model to an agent B of one of three (or more) possible types. We believe such an extension is nontrivial for the following reason. Given only two types

[11] I'm likely to gain a little and potentially lose a lot from revealing my medical history to a random person, whereas I am likely to gain a lot from truthfully reporting my medical history to a doctor.

($t = 0$ or $t = 1$), there's a single mechanism that is differentially private: Randomized Response. In contrast, the space of differentially private mechanisms over three types is far richer. Consider just the following two ϵ-DP mechanisms: $\mathcal{M}_1$ is the exponential mechanism over the three possible outputs and $\mathcal{M}_2$ is the mechanism that adds geometric noise (integral Laplace noise) proportional to $2/\epsilon$ to the input type and then truncates the result back to $\{0, 1, 2\}$. Assuming that ϵ is a very small constant, we have that under the exponential mechanism it holds that for each $i, j \in \{0, 1, 2\}$ we have $\mathbf{Pr}[\mathcal{M}_1(i) = j] \approx \frac{1}{3}$, whereas for the truncated geometric noise mechanism we have that for all $i \in \{0, 1, 2\}$ it holds that $\mathbf{Pr}[\mathcal{M}_2(i) = 0] \approx \frac{1}{2}$ and $\mathbf{Pr}[\mathcal{M}_2(i) = 2] \approx \frac{1}{2}$, while $\mathbf{Pr}[\mathcal{M}_2(i) = 1] \approx \epsilon$. If we were to interpret each mechanism as a strategy for an agent that broadcasts one of three signals, it is intuitively clear that the incentives motivating a behavior similar to $\mathcal{M}_1$ ought to be quite different than those motivating a BNE strategy similar to $\mathcal{M}_2$.

Lastly, it would be interesting to pursue this line of work further, by studying more complex games. In particular, we propose the following scenario, which resembles the standard narrative in differential privacy literature and should provide a complementary approach to the "sensitive surveyor" problem [13, 14, 20, 22, 23]. Suppose that the signal that B sends is not for a type of coupon that gives B an immediate and fixed reward, but rather a response of B to a survey question. That is, suppose B interacts with a benevolent data curator that wishes to learn the distribution of type-0 and type-1 agents in the population and B may benefit from the effect of the curator's analysis. (For example, the data curator may ask people with a certain disease about their exposure to some substance.) In such a case, B's utility is a function of the curator's ability to well approximate the true answer. In addition to the potential gain, there is also potential loss, based on B's concerns about his or her private information being publicly exposed. What formulation of this privacy loss results in B playing according to a Randomized Response strategy? What explicit formulation of privacy loss causes B to truthfully report his or her type knowing that A's data will be published using an ϵ-differentially private mechanism?

APPENDIX

A ADDITIONAL MATERIAL: COUPON GAME WITH PROPER SCORING RULES

A.1 Background: Proper Scoring Rules

Proper scoring rules (see surveys [15, 27]) were devised as a method to elicit experts to report their true prediction as to the probability of an event happening. That is, given a Bernoulli random variable X, we ask an expert to report his or her estimation of $\mu = \mathbf{Pr}[X = 1]$. Given that the expert reports x, we pay him or her $f_1(x)$ if indeed $X = 1$ and pay him or her $f_0(x)$ otherwise. A *proper scoring rule* is a pair of functions (f_0, f_1) such that $\arg\max_x \mathbf{E}_{t \leftarrow X}[f_t(x)] = \mu$, where the maximum is obtained for a unique report. That is, it is in the expert's best interest to report the true prior. It was shown [15, 24] that a pair of twice-differentiable functions (f_0, f_1) give a proper scoring rule iff there exists a convex function g (i.e., $g'' > 0$ on the $[0, 1]$ interval) s.t. $f_0(x) = g(x) - xg'(x)$, $f_1(x) = g(x) + (1 - x)g'(x)$. Using the derivatives of both functions ($f_0'(x) = -xg''(x)$ and $f_1'(x) = (1 - x)g''(x)$), we deduce that f_0 is a strictly decreasing function and f_1 is a strictly increasing function on the $[0, 1]$ interval. And so, given that $X = 1$ w.p. μ, we have that the expected payment for an expert predicting x is

$$F_\mu(x) = (1 - \mu)f_0(x) + \mu f_1(x) = g(x) - (x - \mu)g'(x), \tag{11}$$

which is maximized at $x = \mu$, where $F_\mu(\mu) = g(\mu)$.

Most commonly discussed proper scoring rules are *symmetric* (or label-invariant) proper scoring rules, which are oblivious to that outcomes of X (also referred to as neutral scoring rules in [5]). That is, symmetric scoring rules have the property that for any two Bernoulli random variables

X and X' s.t. $\mathbf{Pr}[X = 1] = \mathbf{Pr}[X' = 0]$, the expected payment for an expert predicting x for X is identical to the payment for an expert predicting $1 - x$ for X'. Such symmetric scoring rules are derived from a convex function g that is symmetric around $\frac{1}{2}$, i.e., $g(x) = g(1 - x)$, and so $g'(x) = -g'(1 - x)$ and $g''(x) = g''(1 - x)$.

Concrete examples of proper scoring rules, such as the quadratic scoring rule, the spherical scoring rule, and the logarithmic scoring rules, are discussed in Section A.2.

A.2 Strategies under Specific Scoring Rules

We now plug in different types of proper and symmetric scoring rules and find what p^* and q^* are in each case. We analyze the game for a value of ρ s.t. the BNE is obtained where neither p^* nor q^* is integral. We also characterize what is the ϵ in A's posterior probability—the value of $\max_b \left\{\ln\left(\frac{\mathbf{Pr}[\hat{t}=t \mid t=b]}{\mathbf{Pr}[\hat{t}=1-t \mid t=b]}\right)\right\}$.

There exist three canonical rules often used in the literature: Quadratic, Spherical, and Logarithmic.

Quadratic Scoring Rule. The quadratic scoring rule is defined by the functions $(f_0(x), f_1(x)) = (2 - 2x^2, 4x - 2x^2)$. The quadratic scoring rule is generated by the convex function $g(x) = x^2 + (1 - x)^2 + 1 = 2 - 2x + 2x^2$. (So, $g'(x) = -2 + 4x$ and $g''(x) = -2$.) Therefore, $(f_0'(x), f_1'(x)) = (-4x, 4(1 - x))$.

Observe that since $g' \in [-2, 2]$, Equation (6) gives that $\rho \in [-2, 2]$ as well. Hence, Equation (6) takes the form

$$\rho = 2 - \frac{4}{1 + \frac{D_0 p}{D_1(1-q)}} \Rightarrow \frac{D_0 p}{D_1(1-q)} = \frac{2+\rho}{2-\rho} \Rightarrow p = \frac{D_1}{D_0}\frac{2+\rho}{2-\rho}(1-q)$$

$$\rho = -2 + \frac{4}{1 + \frac{D_0(1-p)}{D_1 q}} \Rightarrow \frac{D_0(1-p)}{D_1 q} = \frac{2-\rho}{2+\rho} \Rightarrow q = \frac{D_0}{D_1}\frac{2+\rho}{2-\rho}(1-p).$$

So we have

$$p = \frac{D_1}{D_0}\frac{2+\rho}{2-\rho} - \left(\frac{2+\rho}{2-\rho}\right)^2 (1-p) \Rightarrow p = \left(\frac{2+\rho}{2-\rho}\right)\left(\frac{D_0}{D_1} - \frac{2+\rho}{2-\rho}\right) \Big/ \left(1 - \left(\frac{2+\rho}{2-\rho}\right)^2\right),$$

which boils down to

$$p = \frac{2+\rho}{4}\left(\frac{\frac{2+\rho}{2-\rho} - \frac{D_0}{D_1}}{\frac{2+\rho}{2-\rho} - 1}\right) = \frac{2+\rho}{4}\left(\frac{2(1 - \frac{D_0}{D_1}) + \rho(1 + \frac{D_0}{D_1})}{2\rho}\right) = \frac{2+\rho}{4}\left(\frac{1}{2}\left(1 + \frac{D_0}{D_1}\right) - \frac{D_0 - D_1}{\rho D_1}\right).$$

And similarly,

$$q = \frac{D_0}{D_1}\frac{2+\rho}{2-\rho} - \left(\frac{2+\rho}{2-\rho}\right)^2 (1-q) \Rightarrow q = \left(\frac{2+\rho}{2-\rho}\right)\left(\frac{D_1}{D_0} - \frac{2+\rho}{2-\rho}\right) \Big/ \left(1 - \left(\frac{2+\rho}{2-\rho}\right)^2\right),$$

which gives

$$q = \frac{2+\rho}{4}\left(\frac{\frac{2+\rho}{2-\rho} - \frac{D_1}{D_0}}{\frac{2+\rho}{2-\rho} - 1}\right) = \frac{2+\rho}{4}\left(\frac{2(1 - \frac{D_1}{D_0}) + \rho(1 + \frac{D_1}{D_0})}{2\rho}\right) = \frac{2+\rho}{4}\left(\frac{1}{2}\left(1 + \frac{D_1}{D_0}\right) + \frac{D_0 - D_1}{\rho D_1}\right).$$

More importantly, under these p and q values, $y_0 = \frac{2-\rho}{4}$ and $y_1 = \frac{2+\rho}{4}$. So from A's perspective, there is a Randomized Response move here with $e^\epsilon = y_1/y_0$, and hence,

$$\epsilon = \ln\left(\frac{2+\rho}{2-\rho}\right).$$

The expected utility of A is $u_A = g(y_0) = y_0^2 + y_1^2 + 1 = \frac{8+2\rho^2}{16} + 1 = 2 - \frac{1}{2} + \frac{\rho^2}{8}$. This is in comparison to $g(D_1) = 1 + D_0^2 + D_1^2 = 2 - 2D_1 + 2D_1^2 = 2 - 2D_1(1 - D_1) = 2 - 2D_1D_0$. It follows that A prefers the second game (with the coupon) to the first only if $\frac{1}{2} - \frac{\rho^2}{8} < 2D_0D_1$ or $\rho^2 > 4 - 16D_0D_1$. Clearly, with $D_0 = D_1 = \frac{1}{2}$, we have that A prefers the coupon game over the benchmark game.

Spherical scoring rule. The spherical scoring rule is defined by the functions

$$(f_0(x), f_1(x)) = \left(\frac{1-x}{\sqrt{x^2 + (1-x)^2}}, \frac{x}{\sqrt{x^2 + (1-x)^2}}\right),$$

which are generated using $g(x) = \sqrt{x^2 + (1-x)^2}$. (So, $g'(x) = \frac{2x-1}{\sqrt{x^2+(1-x)^2}}$ and $g''(x) = (x^2 + (1 - x)^2)^{-\frac{3}{2}}$.) Therefore, $(f_0'(x), f_1'(x)) = (-x(1 - 2x + 2x^2)^{-\frac{3}{2}}, (1 - x)(1 - 2x + 2x^2)^{-\frac{3}{2}})$.

Using the definition of $g'(x)$, Equation (6) now yields

$$\rho\sqrt{y_0^2 + 1 - 2y_0 + y_0^2} = -2y_0 + 1 \quad \Rightarrow (4 - 2\rho^2)y_0^2 - (4 - 2\rho^2)y_0 + (1 - \rho^2) = 0$$
$$\rho\sqrt{y_1^2 + 1 - 2y_1 + y_1^2} = 2y_1 - 1 \quad \Rightarrow (4 - 2\rho^2)y_1^2 - (4 - 2\rho^2)y_1 + (1 - \rho^2) = 0.$$

So y_0 and y_1 are the two different roots of the equation $x^2 - x + \frac{1-\rho^2}{4-2\rho^2} = 0$, namely $\frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\rho^2}{2-\rho^2}}$. Plugging in the values of y_0 and y_1, we have

$$\frac{D_1q - D_0(1-p)}{D_1q + D_0(1-p)} = \sqrt{\frac{\rho^2}{2-\rho^2}}$$
$$\frac{D_0p - D_1(1-q)}{D_0p + D_1(1-q)} = \sqrt{\frac{\rho^2}{2-\rho^2}},$$

and this is because we assume $D_0p > D_1(1 - q)$ and $D_1q > D_0(1 - p)$. (That is, when we see the signal $\hat{t} = 0$, it is more likely to come from a $t = 0$-type agent than a $t = 1$-agent, and similarly with the $\hat{t} = 1$ signal.)

After arithmetic manipulations, we have

$$(1 - \rho^2)(D_0^2p^2 + D_1^2(1 - q)^2) = 2D_0D_1p(1 - q) \quad \Rightarrow (1 - \rho^2)D_0p = D_1(1 - q)\left(1 \pm \rho\sqrt{2 - \rho^2}\right)$$
$$(1 - \rho^2)(D_0^2(1 - p)^2 + D_1^2q^2) = 2D_0D_1(1 - p)q \quad \Rightarrow (1 - \rho^2)D_1q = D_0(1 - p)\left(1 \pm \rho\sqrt{2 - \rho^2}\right)$$

using the fact that $\rho \le 1$ and that $D_0p > D_1(1 - q)$ and $D_1q > D_0(1 - p)$; then

$$D_0p = D_1(1 - q)\frac{1 + \rho\sqrt{2 - \rho^2}}{1 - \rho^2} \stackrel{\text{def}}{=} Z_\rho D_1(1 - q)$$
$$D_1q = D_0(1 - p)\frac{1 + \rho\sqrt{2 - \rho^2}}{1 - \rho^2} \stackrel{\text{def}}{=} Z_\rho D_0(1 - p)$$

(because $1 - \rho\sqrt{2 - \rho^2} \le 1 - \rho \le 1 - \rho^2$). We have that

$$D_1 = D_1q + D_1(1 - q) = Z_\rho D_0(1 - p) + \frac{1}{Z_\rho}D_0p$$
$$D_0 = D_0p + D_0(1 - p) = Z_\rho D_1(1 - q) + \frac{1}{Z_\rho}D_1q.$$

We deduce

$$p = \frac{Z_\rho^2 - Z_\rho \frac{D_1}{D_0}}{Z_\rho^2 - 1}, \qquad q = \frac{Z_\rho^2 - Z_\rho \frac{D_0}{D_1}}{Z_\rho^2 - 1}.$$

More importantly, from A's perspective, the signal is like a Randomized Response with parameter $e^\epsilon = y_1/y_0$ so

$$\epsilon = \ln\left(\left(1 + \sqrt{\frac{\rho^2}{2-\rho^2}}\right) \Big/ \left(1 - \sqrt{\frac{\rho^2}{2-\rho^2}}\right)\right).$$

The utility of A from the game is now $g(y_0)$, which boils down to $\frac{1}{2-\rho^2}$. This is in contrast to $D_0^2 + D_1^2$, so A prefers the game with the coupon over the baseline when $\rho^2 > 2 - \frac{1}{D_0^2+D_1^2} = \frac{(D_0-D_1)^2}{D_0^2+D_1^2}$. Complementary to that, B's expected payment is

$$\rho(D_0 p + D_1 q) - g(y_0) = \rho\left(\frac{D_0 Z_\rho^2 - D_1 Z\rho + D_1 Z_\rho^2 - D_0 Z_\rho}{Z_\rho^2 - 1}\right) - \frac{1}{2-\rho^2} = \frac{\rho Z_\rho}{Z_\rho + 1} - \frac{2}{2-\rho^2}.$$

Logarithmic scoring rule. The logarithmic scoring rule is defined by the functions $(f_0(x), f_1(x)) = (\ln(1-x), \ln(x))$, which are generated by $g(x) = -H(x) = x\ln(x) + (1-x)\ln(1-x)$. (So, $g'(x) = \ln(x) - \ln(1-x)$ and $g''(x) = \frac{1}{x} + \frac{1}{1-x}$.) Therefore, $(f_0'(x), f_1'(x)) = (-\frac{1}{1-x}, \frac{1}{x})$. Observe that the logarithmic scoring rule has *negative* costs, and furthermore, we may charge infinite cost from an expert reporting $x = 0$ or $x = 1$.

Using $g'(x)$, Equation (6) takes the form

$$\rho = \ln\left(\frac{1-y_0}{y_0}\right) = \ln\left(\frac{y_1}{1-y_1}\right) \;\Rightarrow y_0 = \frac{1}{1+e^\rho}, y_1 = \frac{1}{1+e^{-\rho}}.$$

This implies that

$$\frac{D_0 p}{D_1(1-q)} = \frac{D_1 q}{D_0(1-p)} = e^\rho \;\Rightarrow\; p = \frac{e^{2\rho} - e^\rho \frac{D_1}{D_0}}{e^{2\rho}-1}, q = \frac{e^{2\rho} - e^\rho \frac{D_0}{D_1}}{e^{2\rho}-1}.$$

The Randomized Response behavior that A observes is for $e^\epsilon = y_1/y_0$, which means that simply $\epsilon = \rho$. The utility for A is now $g(y_0) = -\frac{\ln(1+e^\rho)}{1+e^\rho} - \frac{\ln(1+e^{-\rho})}{1+e^{-\rho}}$. And the utility for B is $u_B = \rho(D_0 p + D_1 q) - g(y_0) = \rho\frac{e^{2\rho}-e^\rho}{e^{2\rho}-1} - g(y_0) = \frac{\rho e^\rho}{e^\rho+1} - g(y_0)$.

ACKNOWLEDGMENTS

We would like to thank Kobbi Nissim for many helpful discussions and helping us in initiating this line of work. We thank the anonymous referees for many helpful suggestions that allowed us to significantly improve the presentation of this work.

REFERENCES

[1] Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. 2013. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *Foundations of Computer Science (FOCS'13)*.

[2] Dirk Bergemann, Benjamin Brooks, and Stephen Morris. 2013. *The Limits of Price Discrimination.* Cowles Foundation Discussion Papers 1896. Cowles Foundation for Research in Economics, Yale University. Retrieved from http://ideas.repec.org/p/cwl/cwldpp/1896.html.

[3] Giacomo Calzolari and Alessandro Pavan. 2006. On the optimality of privacy in sequential contracting. *Journal of Economic Theory* 130, 1 (2006), 168–204.

[4] Yiling Chen, Stephen Chong, Ian A. Kash, Tal Moran, and Salil P. Vadhan. 2013. Truthful mechanisms for agents that value privacy. In *EC*.

[5] Yiling Chen, Nikhil R. Devanur, David M. Pennock, and Jennifer Wortman Vaughan. 2014. Removing arbitrage from wagering mechanisms. In *EC*.

[6] Vincent Conitzer, Curtis R. Taylor, and Liad Wagman. 2012. Hide and seek: Costly consumer privacy in a market with repeat purchases. *Marketing Science* 31, 2 (2012), 277–292.
[7] Cynthia Dwork. 2006. Differential privacy. In *ICALP*.
[8] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*.
[9] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*.
[10] Cynthia Dwork and Adam Smith. 2010. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* 1, 2 (2010), 2.
[11] Lisa Fleischer and Yu-Han Lyu. 2012. Approximately optimal auctions for selling privacy when costs are correlated with data. In *EC*.
[12] D. Fudenberg, J. Tirole, J. A. Tirole, and MIT Press. 1991. *Game Theory*. MIT Press.
[13] Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. 2014. Buying private data without verification. In *EC*.
[14] Arpita Ghosh and Aaron Roth. 2011. Selling privacy at auction. In *EC*.
[15] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102, 477 (2007), 359–378.
[16] Ronen Gradwohl and Rann Smorodinsky. 2014. Subjective perception games and privacy. *CoRR* abs/1409.1487 (2014). http://arxiv.org/abs/1409.1487.
[17] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2008. What can we learn privately? In *FOCS*.
[18] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. Oxford University Press.
[19] Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *FOCS*. 94–103.
[20] Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. 2012. Privacy-aware mechanism design. In *EC*.
[21] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. 2012. Approximately optimal mechanism design via differential privacy. In *ITCS*.
[22] Kobbi Nissim, Salil P. Vadhan, and David Xiao. 2014. Redrawing the boundaries on purchasing data from privacy-sensitive individuals. In *ITCS*.
[23] Aaron Roth and Grant Schoenebeck. 2012. Conducting truthful surveys, cheaply. In *EC*.
[24] Leonard J. Savage. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66, 336 (1971), 783–801.
[25] Michael Spence. 1973. Job market signalling. *Quarterly Journal of Economics* 87, 3 (August 1973), 355–374.
[26] Stanley L. Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60, 309 (March 1965), 63–69.
[27] R. L. Winkler. 1996. Scoring rules and the evaluation of probabilities. *Test* 5, 1 (1996), 1–60.
[28] David Xiao. 2013. Is privacy compatible with truthfulness? In *ITCS*. 67–86.

Received February 2019; revised December 2019; accepted January 2020