Privacy in Implementation*

Ronen Gradwohl†

Abstract

In most implementation frameworks, agents care only about the outcome and not at all about the way in which it was obtained. Additionally, typical mechanisms for full implementation involve the complete revelation of all private information to the planner. In this paper I consider the problem of full implementation with agents who may prefer to protect their privacy. I analyze the extent to which privacy-protecting mechanisms can be constructed under various assumptions about agents' predilection for privacy and the permissible game forms.

Keywords: Nash implementation, subgame perfect implementation, privacy.

1 Introduction

The recent privacy-related charges made by the Federal Trade Commission against Facebook and Google are typical examples of individuals' growing concerns about the exposure of their private information. These concerns, however, are not specific to the case of digital privacy and actually arise in many diverse strategic situations. For example, a cabinet member who casts a vote in a cabinet meeting may care about the actual effect of his vote, but if the position he prefers is perceived as unfavorable by his constituents, he may be hesitant to expose his preference. A buyer who negotiates the trade of a good with a seller may prefer to keep his valuation of the good private,

*I gratefully acknowledge NSF award #1216006. I would like to thank Ehud Kalai, Ariel Rubinstein, Yuval Salant, Ron Siegel, and Rakesh Vohra for helpful conversations about this research. I am also grateful to seminar participants at Northwestern University, Tel Aviv University, Hebrew University, Harvard, and MIT for valuable feedback.

†Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA. E-mail: r-gradwohl@kellogg.northwestern.edu.

because revelation of this information may weaken his bargaining position in future dealings with other sellers. An individual who testifies in a court of law or participates in a government hearing may wish to keep certain information private if it casts him in an unfavorable light or causes him some embarrassment. In all these examples, individuals are concerned not only with the final material outcome of their actions, but also with the private information that is revealed in the course of interaction.

The core of this paper is a model in which agents have such *information-sensitive* preferences—they care not only about the outcomes of an interaction, but also about the type and amount of private information that is revealed. There are many reasons agents might care about such information revelation: for example, the information that is potentially revealed may have material consequences in later interactions, but it may be difficult to model all future interactions and determine how they are affected by this information. Additionally, in some situations a planner may only have control over a particular interaction, even if the information that is revealed in that interaction may have material consequences for the agents in the future. So when a planner designs a mechanism for his particular interaction, he must view the agents as having information-sensitive preferences. In any case, such information-sensitive preferences cannot be captured by standard models in economics or game theory.

The goal of this paper is twofold. First, it highlights the subtleties involved in modeling and reasoning about such information-sensitive preferences. And second, it initiates a study of implementation for agents with such preferences—it delineates the line between what is possible and what is impossible, and in particular contrasts this with implementation in the standard setting.

More specifically, in this paper I study the strategic effects of information-sensitive preferences in the context of full implementation[1] with complete information. With complete information, agents already know each other's private information, so all privacy concerns are vis-à-vis the planner and other outside observers. In this framework, the presence of information-sensitive preferences introduces two issues into the theory of implementation: The first is that known mechanisms for full implementation may no longer work, since the set of equilibria under information-sensitive preferences may be different from the set of equilibria under standard preferences.

[1]With full implementation, the desideratum is a mechanism in which *all* equilibria lead to socially optimal outcomes. For various surveys of this vast literature, see Jackson (2001), Maskin and Sjöström (2002), or Palfrey (2002). The case of partial implementation, in which only *one* equilibrium must lead to a socially optimal outcome, is also affected to some degree by the presence of information-sensitive preferences—see Section 6.1.

The second issue arises from the observation that in most known mechanisms for full implementation, all private information is revealed.[2] But if agents prefer to keep their information private, it may be desirable to design mechanisms that preserve privacy to some extent.

In this paper I examine both of the following questions: Are there mechanisms that are *information-sensitive* implementations, in which all equilibria with respect to information-sensitive preferences achieve a social optimum? And are there mechanisms that are *privacy-protecting*, in which some equilibria reveal none of the agents' private information other than what is implied by the outcome itself?

Observe that the first question is not specifically about privacy, but rather about information-sensitive preferences more generally. In particular, this question and the answers that will follow apply also in settings where agents *want* to disclose some or all of their private information. The second question is specific to privacy, and an assumption that will underlie some of the answers is that agents prefer privacy. However, the possibility results for this question can actually be extended to a setting in which agents have more nuanced preferences over what information is revealed and, in particular, to a setting in which they want to reveal some or all of their private information.[3]

A Simple Example To illustrate the difficulties of implementation with information-sensitive preferences, as well as provide an overview of the results of this paper, I describe a simple example with three agents and two alternatives $\{0, 1\}$. Each agent can be one of two types $\{t_0, t_1\}$, where type t_i *intrinsically* prefers outcome i to outcome $1 - i$. This means that type t_i prefers outcome i to $1 - i$ when all private information—namely, the profile of all players' types—is revealed. Furthermore, agents' information-sensitive preferences are such that, for any fixed outcome, they strictly prefer that their true type not be revealed, regardless of the information revealed about other players' types. That is, they prefer both privacy, in which the planner does not learn their true type, and deception, in which the planner incorrectly learns their type, to the true revelation of their type. I will later specify the agents' information-sensitive preferences when the outcome is not fixed.

Suppose a planner wishes to implement the majority of agents' intrinsic preferences. He may design a simple voting mechanism: each agent reports an action that

[2]See, for example, Maskin (1999) and Moore and Repullo (1988).

[3]Section 6.2 discusses this extension.

can be 0 or 1, and the outcome is the majority of the actions. Suppose also that the planner expects agents to vote truthfully—they vote for i if their type is t_i. Now, voting truthfully is an equilibrium with standard preferences, but is it an equilibrium with information-sensitive preferences?

When preferences are information sensitive, equilibria relate to both the outcomes and revealed information as follows. For each profile of types, each action profile corresponds to a pair (a, S), where a is the outcome obtained by the action profile and S is the set of possible types that the planner believes are possible. An equilibrium is then a strategy profile in which the pairs (a, S) obtained in equilibrium are preferred by all agents to pairs (b, T) obtained by unilateral deviations. But where do the sets of possible types come from? Suppose that agents play a strategy profile s, and some action profile is realized. Then the set of possible types consists of all type profiles R that could have led to the realized action profile.

In the majority example, the profile in which agents always vote truthfully leads to full revelation of information, since any realized action profile uniquely identifies all agents' types. However, this is not an equilibrium. To see this, suppose all agents are of type t_0. Then one agent's unilateral deviation will not change the outcome, since the majority will still be 0. However, a unilateral deviation by an agent will lead to a different set of possible types in which the planner incorrectly believes this agent is of type t_1. Agents prefer deception over revelation, and so this deviation is profitable.

An additional problem with the simple voting mechanism is that it is not a full implementation—there are other strategy profiles that are equilibria but that do not yield the majority. A simple example is the profile in which all agents always vote for 0.

To address both of these difficulties, one might turn to more-complex mechanisms. Maskin (1999) shows that under some conditions that are satisfied in the majority example, his mechanism is a full implementation. But is it a full implementation also with information-sensitive preferences? It turns out that under a mild assumption about the planner's beliefs at action profiles that are not reachable in equilibrium, truthfulness is an information-sensitive equilibrium of this mechanism.[4] However,

[4]The difficulty is that a unilateral deviation may lead to an action profile that is not reachable in equilibrium, and so the planner may be unable to invert this off-equilibrium profile and derive the possible types. The mild assumption—called *one-deviation consistency*—is roughly the following: For an off-equilibrium action profile that is "close" to the equilibrium path in the sense that it is reachable by type profiles belonging to a set U but allowing for a unilateral deviation, the set of

whether or not there are other, undesirable equilibria depends on some aspects of agents' information-sensitive preferences that I have not yet specified. While agents prefer privacy over revelation when the outcome is *fixed*, what is their preferences when it is not? In particular, does type t_i prefer outcome i with revelation of information or outcome $1-i$ with privacy? If he prefers the latter, then full implementation is impossible, and there will always be equilibria that do not yield the majority. This is formalized in Proposition 3.1, which shows that without restrictions on information-sensitive preferences, implementation by *any* mechanism may be impossible. If type t_i prefers outcome i with revelation, however, then the mechanism of Maskin (1999) is a full implementation. This is a consequence of Proposition 5.1. In fact, that proposition shows that if agents' preferences are such that they are willing to reveal their private information for some better outcome, then full implementation is possible with information-sensitive preferences *whenever* it is possible with standard preferences.

Even if the mechanism of Maskin (1999) works, however, there is still a problem: In that mechanism, agents must reveal all their private information. But do there exist mechanisms for full implementation of majority that do not reveal information beyond the outcome? Theorem 4.2 shows that this is impossible. Furthermore, under some conditions, the unique non-constant social-choice function that can be implemented while protecting privacy is the dictatorship function.

However, not all is lost. If we allow for extensive-form mechanisms, in which communication proceeds in stages, then such privacy-protecting implementation of majority becomes possible. The possibility result is quite strong: Theorem 5.2 shows that with extensive-form mechanisms, privacy-protecting implementation is possible *whenever* implementation with standard preferences is possible.

Related Literature While in standard economic and game theoretic models agents have preferences only over material outcomes, there are two strands of the literature that consider agents who care also about the private information that is revealed. The literature on social image, including Bernheim (1994), Glazer and Konrad (1996), and Ireland (1994), studies the behavioral effects of agents' concerns for how they are perceived by others. The more general literature on psychological games studies agents who, in addition to being concerned with physical outcomes, also care about their

possible types can be any nonempty subset of U. For all other action profiles, the set of possible types is unrestricted. See Section 2.4 for details.

beliefs and the beliefs of others (Geanakoplos et al. (1989)). However, while the modeling of agents in these literatures is related to this paper's model of agents with information-sensitive preferences, their aims are very different. In particular, these areas of research are typically not concerned with the problem of designing mechanisms for such agents, but rather with the study of various behavioral phenomena resulting from such agents' preferences.

Preferences that are not only over outcomes do appear in two implementation frameworks. Glazer and Rubinstein (1998) study an implementation problem in which agents may be motivated not only by the material outcomes of their actions, but also by the desire to have their own recommendation accepted. Matsushima (2008a,b) and Dutta and Sen (2012) study the problem of full implementation when agents have a strict preference for *honesty* when the resulting material outcome is not worsened.

The desire to control information transmission is also present endogenously and implicitly in many models of repeated interactions. In such interactions, agents play a strategy that balances between myopic payoff maximization and information revelation that may potentially lead to a lower payoff in the future. This theme is prevalent in the literature on dynamic mechanism design (c.f. Vohra (2012); Pavan et al. (2012)). The current paper is related in that it is also partly motivated by agents' concerns about revealing information that may impact later interactions. However, it is different in that it models such concerns exogenously, taking the position that it is often both impractical and unrealistic to model all future interactions. In fact, realistically it often seems difficult to justify a model in which both the designer and the agents have complete knowledge of all future interactions.

Also related is work by Calzolari and Pavan (2006b,a), who examine the problem of the optimal information disclosure policy in two-stage interactions. In some settings they show that privacy is, in fact, an optimal policy for the principal. However, while these papers do study the interplay between information disclosure and economic interaction, the preference for privacy is solely the principal's—the agents are assumed to have standard preferences over outcomes.

Finally, this paper is related to a vast literature on privacy in computer science, particularly to privacy concerns in cryptography and to the newer study of differential privacy (Dinur and Nissim (2003)). These have been applied to strategic settings: for example, Naor et al. (1999) design a cryptographic system for guaranteeing privacy in auctions, and McSherry and Talwar (2007) utilize the tools of differential privacy to design economic mechanisms. However, in these applications, agents are not modeled

as caring about privacy. That is, all these applications achieve some goal privately, but only when agents do not care about privacy. Once agents have preferences that depend on the information that is revealed, these applications can break down.

A number of more recent works do model agents' predilection for privacy explicitly by including a cost incurred by agents when some or all of their information is revealed. Miltersen et al. (2009) focus on the cryptographic implementation of a first-price auction when agents have marginal privacy concerns. Ghosh and Roth (2011) study the design of markets for selling privacy. Xiao (2011) studies the question of whether differential privacy is sufficient for truthful revelation of private information. Nissim et al. (2012) and Chen et al. (2011) consider a more general mechanism design problem in which agents are concerned about the information leaked by the outcome of a mechanism, and where all communication is hidden by perfect cryptography or a trusted third party.

While the motivation for these papers is similar to some of the motivation underlying the current paper, they are quite distinct. These papers all assume that communication is perfectly hidden by cryptography, and there are many situations in which the use of such technology is not feasible. For example, the US Senate and House of Representatives conduct many recorded votes each year, in which the votes of all participants are publicized. When a buyer makes an online purchase, the vendor is generally aware of this purchase. When an individual testifies in a court of law, his testimony is heard by all present. In such cases many of the cryptographic tools are not applicable.

Furthermore, Nissim et al. (2012) and Chen et al. (2011) utilize tools from differential privacy to obtain mechanisms that do not reveal much of the agents' private information. There are two inherent features of such tools that make them inapplicable in many settings. First, they only apply when the number of agents is very large. Second, these mechanisms are randomized and do not always correctly implement the social choice function. In this sense they are close to the notion of virtual implementation of Abreu and Sen (1991). Unlike that notion, however, the approximation obtained is not arbitrarily close, but rather becomes close only as the number of agents becomes large. For a fixed number of agents, the mechanisms of Nissim et al. (2012) and Chen et al. (2011) could yield a suboptimal outcome with non-negligible probability.

Finally, in terms of the setting for implementation, the works of Nissim et al. (2012) and Chen et al. (2011) are incomparable to the current paper. The former

consider partial implementation in dominant strategies in a setting with incomplete information and cardinal preferences, whereas the current paper examines full implementation in Nash and subgame perfect equilibrium in a setting with complete information and ordinal preferences.

Organization The rest of the paper is organized as follows. Section 2 presents the model. Sections 3 and 4 contain impossibility results—in the former I show that without restrictions on agents' privacy concerns there can be no implementation, and in the latter I show that, even with restrictions, almost nothing can be implemented in a privacy-protecting manner with normal-form mechanisms. Section 5 contains possibility results on information-sensitive and privacy-protecting implementation, and Section 6 contains further extensions of the model and results. Finally, the Appendix contains all proofs that do not appear in the main body of the text.

2 The Model

N denotes a finite set of agents and also, with a slight abuse of notation, its cardinality. $\mathcal{O}$ denotes a possibly infinite set of outcomes.

2.1 Preferences and Social Choice Correspondences

We begin with the usual setup of agent preferences, but because these standard preferences will be extended we refer to them as the *intrinsic* preferences. The intrinsic preferences of an agent i are represented by a complete, transitive, binary relation R_i over $\mathcal{O}$, where aR_ib if agent i weakly prefers outcome $a \in \mathcal{O}$ over outcome $b \in \mathcal{O}$. Strict preference is denoted by P_i, and $\text{TR}_i(R) = \{a \in \mathcal{O} : aR_ib \ \forall b \in \mathcal{O}\}$ is the set of top-ranked alternatives for i under R. Denote by $R = (R_1, \ldots, R_N)$ an intrinsic preference profile, and by $(R_{-i}, \overline{R}_i)$ the intrinsic preference profile R in which agent i's intrinsic preferences are replaced by $\overline{R}_i$. Finally, denote by $\mathcal{R}$ the set of admissible profiles of intrinsic preferences.

A social choice correspondence (SCC) $F : \mathcal{R} \twoheadrightarrow \mathcal{O}$ is a mapping from a profile R of intrinsic preferences to a set of outcomes. An SCC is called a social choice function (SCF) if the range is always a singleton, i.e., if $F : \mathcal{R} \mapsto \mathcal{O}$. In this case denote the function by a lower case f. Denote by $F(\mathcal{R})$ the range of F when the domain is $\mathcal{R}$, namely $F(\mathcal{R}) = \cup_{R \in \mathcal{R}} F(R)$. Finally, an SCC F is *constant* if there exists some $a \in \mathcal{O}$

such that $a \in F(R)$ for all $R \in \mathcal{R}$, and otherwise F is *non-constant*.

For a given SCC F and an outcome $a \in \mathcal{O}$, denote by $\mathcal{R}|_{a,F} = \{R \in \mathcal{R} : a \in F(R)\}$. This is the set of profile preferences for which a is a possible outcome under F. When F is clear from context I will simply denote this set as $\mathcal{R}|_a$.

In this paper I extend preferences by adding privacy concerns for the agents. Agents' information-sensitive preferences depend not only on R, but also on a privacy state ψ. More formally, each agent i has preferences R_i^ψ that extend his intrinsic preferences R_i, where R_i^ψ is a complete, transitive, binary relation over $\mathcal{O} \times 2^{\mathcal{R}}$. The first coordinate is an element of $\mathcal{O}$, the outcome, and the second coordinate is a subset of $\mathcal{R}$, which I call the *set of possible types*. The set of possible types is the set of *intrinsic* preference profiles that an outside observer (such as the planner) believes are possible. For example, if a run of a mechanism reveals no information about the intrinsic preferences, then the set of possible types is all of $\mathcal{R}$. If a run of a mechanism only reveals that the true intrinsic preferences R are such that a social choice function f satisfies $f(R) = a$, then the set of possible types is $\mathcal{R}|_a$. If a run of a mechanism reveals the true intrinsic preferences to be some $R \in \mathcal{R}$, then the set of possible types is $\{R\}$.

Now, for any agent i, any $S, T \subseteq \mathcal{R}$, and any $a, b \in \mathcal{O}$, it holds that $(a, S)\ R_i^\psi\ (b, T)$ if and only if agent i weakly prefers outcome a and set of possible types S over outcome b and set of possible types T. Denote by P_i^ψ the strict part of R_i^ψ.

The relation between information-sensitive preferences of an agent and his intrinsic preferences is that the latter are his preferences over outcomes when all intrinsic information is revealed. Formally, I will write the intrinsic preferences as aR_ib or $aR_i^\psi b$, and this will be equivalent to $(a, \{R\})\ R_i^\psi\ (b, \{R\})$. An implication of this is the following. Denote the set of all admissible privacy states by Ψ. Now, for any $i \in N$, any $\psi, \psi' \in \Psi$ and any $R \in \mathcal{R}$, it should be the case that $(a, \{R\})\ R_i^\psi\ (b, \{R\})$ if and only if $(a, \{R\})\ R_i^{\psi'}\ (b, \{R\})$. That is, the intrinsic preferences R_i completely determine the agent's preferences when all information about intrinsic preferences is revealed, and the privacy state ψ extends these preferences to the full domain $\mathcal{O} \times 2^{\mathcal{R}}$.

Observe that this extended framework is a strict generalization of the standard framework, since one can model agents as not caring about privacy. Any agent i can express an unconditional preference of outcome a over b by preferring the pair (a, S) over (b, T) for all sets S and T. Denote by o the privacy state in which this is the case. Formally, for every $i \in N$, every $R \in \mathcal{R}$, every $S, T \subseteq \mathcal{R}$, and any $a, b \in \mathcal{O}$, it holds that $(a, S)R_i^o(b, T)$ if and only if $aR_i^o b$.

In the remainder of the paper I will sometimes refer to R^ψ as the state.

2.2 Mechanisms

A mechanism is an extensive-form game, together with a mapping from terminal histories to elements of $\mathcal{O}$. Formally:

Definition 2.1 (mechanism) *An N-person* mechanism *is a tuple (H, A, g) where*

- *H is a set of (finite)* history *sequences such that the empty word $\epsilon \in H$. A history $h \in H$ is* terminal *if $\{a : (h, a) \in H\} = \emptyset$. The set of terminal histories is denoted Z.*
- *$A = (A_1, \ldots, A_N)$, where each A_i is a function that, for every non-terminal history $h \in H \setminus Z$, assigns a set $A_i(h)$ of actions available to agent i, where $(h, a) \in H$ for all $a \in A_1(h) \times \ldots \times A_N(h)$.*
- *$g : Z \mapsto \mathcal{O}$ is a function that maps terminal histories to outcomes.*

A mechanism is called a normal-form mechanism *if all non-empty histories are terminal.*

2.3 Strategies

A strategy s_i of agent i in a mechanism (H, A, g) is a function that maps any pair (R^ψ, h) to an action from $A_i(h)$. Denote by $s = (s_1, \ldots, s_n)$ a profile of strategies, by $s(R^\psi)$ the profile of strategies in state R^ψ, and by $H(s(R^\psi))$ the terminal history reached when the profile s is played in state R^ψ.

An R^ψ-deviation from a strategy s_i by an agent i is a strategy s_i' that agrees with s_i in every state except R^ψ. Formally, $s_i'(\overline{R}^\psi, h) = s_i(\overline{R}^\psi, h)$ for all h and $\overline{R}^\psi \neq R^\psi$. In state R^ψ the strategy s_i' may be different from s_i. Denote by $s_{-i} \circ s_i'$ the strategy profile in which agent i plays s_i' and agents $j \neq i$ play s_j. Also, denote by $s(R^\psi)|_h$ the profile of strategies s in the subgame rooted at h when the preference profile is R^ψ, and by $H(s(R^\psi)|_h)$ the terminal history reached with this profile in the subgame rooted at h.

2.4 Equilibria

We begin with an informal description of equilibria in normal-form mechanisms. This is followed by formal definitions for both normal-form and extensive-form notions.

In the usual setup, each pure strategy profile corresponds to an outcome dictated by the mechanism. A (pure) Nash equilibrium is then a profile in which no agent has a profitable unilateral deviation. In our setup, however, each strategy profile will correspond to a pair—an outcome a and a set of possible types S. An equilibrium will then be a strategy profile in which no agent can unilaterally deviate to obtain a more favorable pair (b, T).

The outcomes a and b are determined by the function g of the mechanism. But where do the sets of possible types S and T come from? The following is motivated by the notion of a perfect Bayesian equilibrium. Suppose that in some mechanism agents play a strategy profile s, that the privacy state is some ψ, and that some action profile is realized. Then the set of possible types at this action profile is

$$\left\{R : s(R^\psi) \text{ leads to the realized action profile}\right\}.$$

That is, on the path of a profile s, the set of possible types is precisely the set of preferences that lead to the realized action profile.

Next, what is the set of possible types at action profiles that cannot be reached by the strategy profile s? This question is similar to the question of off-equilibrium beliefs in perfect Bayesian equilibrium. In this paper I will typically make a weak restriction on these "off-equilibrium beliefs," called *one-deviation consistency*. Roughly, one-deviation consistency requires that if an action profile cannot be reached by s, but can be reached by a unilateral deviation from s, then the set of possible types is a nonempty subset of all the states from which a unilateral deviation from s leads to that action profile. One-deviation consistency does not at all restrict the set of possible types at histories that cannot be reached by s or by a unilateral deviation from s.

To formalize the discussion above fix a normal-form mechanism (H, A, g) and suppose that the privacy state is some ψ, that the agents play a strategy profile s, and that the terminal node reached is some $z \in Z$. Then define the sets

$$\mathrm{L}^\psi(z, s) = \{R \in \mathcal{R} : H(s(R^\psi)) = z\}$$

and

$$\mathrm{LD}^\psi(z, s) = \{R \in \mathcal{R} : H(s_{-i} \circ s_i'(R^\psi)) = z \text{ for some } i \in N \text{ and strategy } s_i' \text{ of agent } i\}.$$

Note that both L and LD may be empty for terminal histories that cannot be reached by s or a deviation from s. The sets L and LD characterize the preferences that are feasible under the restrictions that agents play a known strategy profile or when only one agent deviates.

Next, we define the set of possible types PT. For each z, s, and ψ define $\mathrm{PT}^{\psi}(z,s)$ to be a subset of $\mathcal{R}$ that satisfies the following:

(i) If $\mathrm{L}^{\psi}(z,s) \neq \emptyset$ then $\mathrm{PT}^{\psi}(z,s) = \mathrm{L}^{\psi}(z,s)$.

The set of possible types PT is *one-deviation consistent* at ψ if the following is also satisfied:

(ii) For all z and s, if $\mathrm{LD}^{\psi}(z,s) \neq \emptyset$, then $\mathrm{PT}^{\psi}(z,s) \neq \emptyset$ and $\mathrm{PT}^{\psi}(z,s) \subseteq \mathrm{LD}^{\psi}(z,s)$.

Note that PT sets that are one-deviation consistent always exist—in particular, one can take $\mathrm{PT}^{\psi}(z,s) = \mathrm{L}^{\psi}(z,s)$ for all s, ψ, and z on the equilibrium path, and $\mathrm{PT}^{\psi}(z,s) = \mathrm{LD}^{\psi}(z,s)$ for z off the equilibrium path.

We now define one notion of equilibrium that we will use—a variant of Nash equilibrium, broadened to allow for preferences over both outcomes and sets of possible types.

Definition 2.2 (information-sensitive Nash equilibrium) *A profile of strategies s in a mechanism (H, A, g) is an* information-sensitive Nash equilibrium *at ψ if for every $i \in N$, $R \in \mathcal{R}$, R^{ψ}-deviations s_i' of agent i, and all sets* PT *that are one-deviation consistent at ψ*

$$\left(g(z), \mathrm{PT}^{\psi}(z,s)\right) \; R_i^{\psi} \; \left(g(z'), \mathrm{PT}^{\psi}(z',s)\right),$$

where $z = H(s(R^{\psi}))$ and $z' = H(s_{-i} \circ s_i'(R^{\psi}))$.

We now define the subgame perfect variant of Definition 2.2. For this, suppose the mechanism (H, A, g) is not a normal-form mechanism. We first extend our restrictions on PT to a dynamic setting by defining sets PT_h for every nonterminal $h \in H$.

Suppose that the privacy state is some ψ, that the agents play a strategy profile s, and that the terminal node reached is some $z \in Z$. Then define the sets

$$\mathrm{L}_h^{\psi}(z,s) = \{R \in \mathcal{R} : H\left(s(R^{\psi})|_h\right) = z\}$$

and

$$\mathrm{LD}_h^{\psi}(z,s) = \{R \in \mathcal{R} : H\left(s_{-i} \circ s_i'(R^{\psi})|_h\right) = z \text{ for some } i \in N \text{ and strategy } s_i' \text{ of agent } i\}.$$

Next, we define the set of possible types PT. For each z, s, ψ, and h, the set $\mathrm{PT}_h^\psi(z,s)$ is a subset of $\mathcal{R}$ that satisfies the following:

(i) If $\mathrm{L}_h^\psi(z,s) \neq \emptyset$ then $\mathrm{PT}_h^\psi(z,s) = \mathrm{L}_h(z,s)$.

The set of possible types PT is *one-deviation consistent* at ψ if the following is also satisfied:

(ii) For all z, s, and h, if $\mathrm{LD}_h^\psi(z,s) \neq \emptyset$, then $\mathrm{PT}_h^\psi(z,s) \neq \emptyset$ and $\mathrm{PT}_h^\psi(z,s) \subseteq \mathrm{LD}_h(z,s)$.

We now define the second notion of equilibrium that we will use. Observe that this definition, when applied to a normal-form mechanism, is equivalent to Definition 2.2.

Definition 2.3 (information-sensitive subgame perfect equilibrium) *A profile of strategies s in a mechanism (H,A,g) is an* information-sensitive subgame perfect equilibrium *at ψ if for every $i \in N$, $R \in \mathcal{R}$, nonterminal $h \in H$, R^ψ-deviations s_i' of agent i, and all sets* PT *that are one-deviation consistent at ψ*

$$\left(g(z), \mathrm{PT}_h^\psi(z,s)\right) \; R_i^\psi \; \left(g(z'), \mathrm{PT}_h^\psi(z',s)\right),$$

where $z = H(s(R^\psi)|_h)$ and $z' = H(s_{-i} \circ s_i'(R^\psi)|_h)$.

The relation between Definition 2.3 and the standard notion of a subgame perfect equilibrium (SPE) is the following: a profile s is an information-sensitive subgame perfect equilibrium at privacy state o if and only if for every profile of preferences $R \in \mathcal{R}$, the strategy profile $s(R^o)$ is a (standard) SPE.

2.5 Implementation

The standard setting A mechanism (H,A,g) is a subgame perfect implementation of an SCC F if for every $R \in \mathcal{R}$ the set of outcomes obtained by subgame perfect equilibria of (H,A,g) is equivalent to $F(R)$.

Abreu and Sen (1990) show that the following condition is necessary for implementation in SPE:

Definition 2.4 (Condition α) *An SCC F satisfies* Condition α *if for all $R, \overline{R} \in \mathcal{R}$ and outcomes $a \in F(R) - F(\overline{R})$ there exist a sequence of agents $j(0), \ldots, j(\ell)$ and a sequence of outcomes $a = a_0, a_1, \ldots, a_\ell, a_{\ell+1}$ such that*

(i) $a_k R_{j(k)} a_{k+1}; k = 0, \ldots, \ell$

(ii) $a_{\ell+1} \overline{P}_j(\ell) a_\ell$

(iii) $a_k \notin \mathrm{TR}_{j(k)}(\overline{R})$ *for* $k = 0, \ldots, \ell$

(iv) if $a_{\ell+1} \in \mathrm{TR}_i(\overline{R})$ *for all* $i \neq j(\ell)$*, then either* $\ell = 0$ *or* $j(\ell - 1) \neq j(\ell)$*.*

Abreu and Sen (1990) also show that Condition α, together with no veto power, is sufficient for subgame perfect implementation[5].

Definition 2.5 (no veto power (NVP)) *An SCC* F *satisfies* no veto power (NVP) *if the following holds for every* $R \in \mathcal{R}$ *and* $a \in \mathcal{O}$*: if* $a \in \mathrm{TR}_i(R)$ *for at least* $N - 1$ *agents* i*, then* $a \in F(R)$*.*

Implementation with privacy With privacy concerns, our definition of information-sensitive implementation of an SCF[6] f (Definition 2.6 below) states that for every R^ψ, the outcome obtained by information-sensitive subgame perfect equilibria should always be $f(R)$. In order to facilitate the modification to privacy-protecting implementation and a strengthening of the definition in Section 6.3, however, we split the definition into parts. Observe that Definition 2.6 is identical to subgame perfect implementation in the standard setting when the privacy state $\psi = o$.

Definition 2.6 (information-sensitive implementation) *A mechanism* (H, A, g) *is an* information-sensitive subgame perfect implementation *of an SCF* f *at* ψ *if:*

1. *There exists a strategy profile* s^* *for which the following hold:*

   (a) s^* *is an information-sensitive subgame perfect equilibrium at* ψ*, and*

   (b) $g(H(s^*(R^\psi))) = f(R)$ *for all* $R \in \mathcal{R}$*.*

2. *For all* $R \in \mathcal{R}$ *and strategy profiles* s *that form an information-sensitive subgame perfect equilibrium at* ψ*, it holds that* $g(H(s(R^\psi))) = f(R)$*.*

[5] Vartiainen (2007) gives conditions for subgame perfect implementation that are both necessary and sufficient.

[6] I focus on implementations of SCFs and not SCCs for simplicity. In Section 6.5 I extend this discussion to SCCs.

Observe that in bullet 2 of Definition 2.6 it is implicitly assumed that the planner knows which strategy profile is being played. This is implicit in the fact that s forms an information-sensitive subgame perfect equilibrium, which, by definition, involves the sets PT that are derived from the profile being played and the observed history. This may be a strong assumption in some settings, and so in Section 6.3 I provide a stronger definition of implementation that dispenses with this assumption.

Next, since we wish to design mechanisms that also protect agents' privacy, we add the following element to our implementations.

Definition 2.7 (privacy-protecting implementation) *A mechanism (H, A, g) is a* privacy-protecting implementation *of an SCF f if it is an information-sensitive implementation of f, and if the strategy profile s^* guaranteed in Definition 2.6 also satisfies the following:*

1. *(c) For all $R, \overline{R} \in \mathcal{R}$ such that $f(R) = f(\overline{R})$, it holds that $H(s^*(\mathcal{R}^\psi)) = H(s^*(\overline{R}^\psi))$.*

Condition 1(c) states that the terminal history $H(s^*(R^\psi))$ could have been reached by s^* with any profile of preferences $\overline{R}^\psi$ for which $f(\overline{R}) = f(R)$. Thus, the planner or any outside observer who sees the outcome $H(s^*(R^\psi)$ cannot differentiate between the true intrinsic preferences being R or $\overline{R}$.

3 Restrictions on information-sensitive Preferences

This section discusses restrictions on information-sensitive preferences. I first show that without any restrictions there may be no information-sensitive implementation.

Proposition 3.1 *Fix any set $\mathcal{R}$ of intrinsic preferences. For each $i \in N$ and $R \in \mathcal{R}$, let R^ψ satisfy the following:*

(i) For any $a, b \in \mathcal{O}$, it holds that $(a, \mathcal{R}) R_i^\psi (b, \mathcal{R})$ if and only if aR_ib.

(ii) For any $a, b \in \mathcal{O}$ and set $S \subsetneq \mathcal{R}$, it holds that $(a, \mathcal{R}) P_i^\psi (b, S)$.

Then at ψ there does not exist an information-sensitive subgame perfect implementation of any non-constant SCF.

The preferences of agents in the proposition are the same as their intrinsic preferences when no information is revealed. The preferences may also be the same as the intrinsic preferences when some information is revealed. However, the key difficulty with these preferences is that regardless of the outcome, each agent strictly prefers *no* information to be revealed than *some* information to be revealed.

Restrictions on information-sensitive preferences Because of the impossibility of implementation with information-sensitive preferences implied by Proposition 3.1, we will restrict the preferences of agents. We will consider two main restrictions. The first is that of lexicographic preferences: roughly speaking, preferences are lexicographic if agents care about privacy only insofar as the outcomes are unaffected. In other words, agents are willing to forego all privacy if they can obtain a more favorable outcome. However, if the outcome is unaffected, they may prefer to reveal as little information as possible.

Definition 3.2 (lexicographic preferences) *ψ is* lexicographic *if for any $i \in N$, $R \in \mathcal{R}$, and sets $S, T \subseteq \mathcal{R}$, it holds that $(a, S) P_i^{\psi} (b, T)$ whenever $a P_i b$.*

The second restriction is that of minimal willingness to reveal (MWR). Roughly speaking, preferences satisfy MWR if agents are willing to reveal all their information for *some* outcome, and particularly for any top-ranked outcome.

Definition 3.3 (MWR preferences) *ψ satisfies* minimal willingness to reveal (MWR) *if for any $i \in N$, $R \in \mathcal{R}$, and sets $S, T \subseteq \mathcal{R}$, it holds that $(a, S) P_i^{\psi} (b, T)$ whenever both $a \in \mathrm{TR}_i(R)$ and $b \notin \mathrm{TR}_i(R)$.*

For implementations that are privacy-protecting and not only information-sensitive, we will need one more restriction on the preferences of agents. This restriction essentially states that for any outcome that is implemented, agents weakly prefer full privacy over full revelation of information.

Definition 3.4 (privacy favoring) *ψ is* privacy favoring *with respect to an SCF f if for each $i \in N$ and $R \in \mathcal{R}$ it holds that*

$$(a, \mathcal{R}|_{a,f}) \; R_i^{\psi} \; (a, \{R\}),$$

where $a = f(R)$.

4 Implementation with Normal-Form Mechanisms

Maskin (1999) shows that Maskin monotonicity (see Definition 4.4) is necessary for implementation with normal-form mechanisms. As this condition is known to be quite restrictive (c.f. Muller and Satterthwaite (1977); Saijo (1987)), there are various relaxations of the problem that allow the theory to be more widely applicable. In particular, the relaxations include restricting the domain of preferences, considering a binary outcome space, and allowing for SCCs rather than SCFs (Postlewaite and Wettstein (1989); Serrano (2004)). In this section I show that even if we allow the first two relaxations, but do require some privacy, then implementation is once again nearly impossible. In Section 6.5 I extend the impossibility results allowing even the third relaxation.

I demonstrate the impossibility of privacy-protecting implementation with normal-form mechanisms via two theorems. The first is the following:

Theorem 4.1 *Fix any domain $\mathcal{R}$ with $|\mathcal{R}| \geq 3$ and any SCF $f : \mathcal{R} \mapsto \{0, 1\}$. Then there exists a lexicographic ψ such that there is no normal-form mechanism that is a privacy-protecting implementation of f at ψ.*

The second impossibility result shows that even if agents do not care about privacy (i.e., $\psi = o$), privacy-protecting implementation is impossible. But first some definitions: A domain $\mathcal{R}$ is *pairwise rich* if for every distinct $i, j \in N$ and $a \neq b \in \mathcal{O}$ there exists some $R \in \mathcal{R}$ such that aP_ib but bP_ja. Also, an SCF f is a *dictatorship* if there exists an agent i such that $f(R) \in \mathrm{TR}_i(R)$ for all $R \in \mathcal{R}$.

Theorem 4.2 *For any pairwise-rich $\mathcal{R}$, if there exists a normal-form mechanism that is a privacy-protecting Nash implementation of an SCF $f : \mathcal{R} \mapsto \{0, 1\}$ at $\mathcal{R}^o$, then f is either constant or a dictatorship.*

Theorem 4.2 is a corollary of Theorem 6.14 from Section 6.5, which essentially states the same for SCCs. The theorem is proved by combining Lemmas 4.5 and 4.6 below, and uses the following definition:

Definition 4.3 (outcome monotonicity (OM)) *An SCF f satisfies* outcome monotonicity (OM) *if for any $\overline{R} \in \mathcal{R}$ and outcome $a \neq f(\overline{R})$, there exists an agent $i \in N$ and an outcome $b \in \mathcal{O}$ for which $b\overline{P}_ia$ but aR_ib for all $R \in \mathcal{R}$ satisfying $a = f(R)$.*

This definition is similar to the well-known definition of Maskin monotonicity:

Definition 4.4 (Maskin monotonicity (MM)) *An SCF f satisfies* Maskin monotonicity (MM) *if for any $R, \overline{R} \in \mathcal{R}$ and outcome a satisfying $a = f(R)$ and $a \neq f(\overline{R})$, there exists an agent $i \in N$ and an outcome $b \in \mathcal{O}$ for which $b\overline{P}_i a$ but $aR_i b$.*

Observe that OM is strictly stronger than MM: OM requires a preference reversal for all R satisfying $a = f(R)$, which in particular implies this reversal for *some* R. On the other hand, there are SCFs that satisfy MM but not OM: the Maj function with strict preferences is such a function.

The two lemmas used in the proof of Theorem 4.2 are the following:

Lemma 4.5 *If there exists a normal-form mechanism that is a privacy-protecting implementation of an SCF f at $\mathcal{R}^o$, then f satisfies OM.*

Lemma 4.6 *For any pairwise-rich $\mathcal{R}$, a non-constant SCF $f : \mathcal{R} \mapsto \{0, 1\}$ satisfies outcome monotonicity if and only if it is a dictatorship.*

5 Implementation with Extensive-Form Mechanisms

In this section I show that information-sensitive and privacy-protecting implementation are possible using extensive-form mechanisms. For information-sensitive implementation I have the following proposition:

Proposition 5.1 *If there is a subgame perfect implementation of an SCF $f : \mathcal{R} \mapsto \mathcal{O}$ that satisfies NVP and $N \geq 3$, then there is an information-sensitive implementation of f for any lexicographic ψ.*

I then extend this proposition and prove the following theorem on the possibility of privacy-protecting implementation:

Theorem 5.2 *If there is a subgame perfect implementation of an SCF $f : \mathcal{R} \mapsto \mathcal{O}$ that satisfies NVP and $N \geq 3$, then there is a privacy-protecting implementation of f for any lexicographic, privacy favoring ψ.*

The proof of Theorem 5.2 is constructive—it describes a mechanism that is a privacy-protecting implementation of f. While the mechanism is a bit intricate, the main idea of the construction is the following. In the first stage of the mechanism, agents attempt to coordinate on an outcome only, without revealing any additional

private information. If there is no unanimous agreement, then the mechanism proceeds with a "contingency plan"—essentially, this is the information-sensitive implementation from Proposition 5.1, in which there is full information revelation. The revelation of information here acts as a sort of "threat" against mis-coordination, and in equilibrium this contingency plan is never invoked. If agents coordinate on an incorrect outcome in the first stage, however, then there will be some agent who will gain by deviating despite the full revelation of information that this will entail.

Proposition 5.1 and Theorem 5.2 provide possibility results for lexicographic ψ. A necessary condition for these implementations is that there be an SPE implementation of f. As Abreu and Sen (1990) show, a necessary condition for SPE implementation is Condition α (see Definition 2.4). For implementation with MWR preferences, as opposed to just lexicographic preferences, we need the following stronger condition:

Definition 5.3 (Condition $\alpha^{\max}$) *An SCC F satisfies* Condition $\alpha^{\max}$ *if it satisfies Condition α but with $(ii)'$ replacing (ii):*

(ii)' $a_{\ell+1} \in \mathrm{TR}_{j(\ell)}(\overline{R})$.

Condition α differs from Condition $\alpha^{\max}$ in item (ii)—the former requires a preference reversal, whereas the latter requires a preference reversal where one outcome becomes top ranked.

The following proposition and theorem are the counterparts to Proposition 5.1 and Theorem 5.2 for MWR preferences:

Proposition 5.4 *If $N \geq 3$ and the SCF $f : \mathcal{R} \mapsto \mathcal{O}$ satisfies Condition $\alpha^{\max}$ and NVP, then there is an information-sensitive implementation of f for any MWR ψ.*

Theorem 5.5 *If $N \geq 3$ and the SCF $f : \mathcal{R} \mapsto \mathcal{O}$ satisfies $\alpha^{\max}$ and NVP, then there is a privacy-protecting implementation of f for any MWR, privacy favoring ψ.*

Remark 5.6 The mechanisms used for the theorems in this section utilize integer games, which are quite unrealistic in real-life mechanisms. Note, however, that the use of such games in our setting is inherited from the mechanisms of Maskin (1999) and Moore and Repullo (1988) which this paper builds on, and without them the results here would probably not be so general. The design of general mechanisms that do not use such games is an open question even in the standard (no-privacy) setting, and is a research agenda orthogonal to this paper. However, I believe that, as in the works of Maskin (1999) and Moore and Repullo (1988), the ideas used in our mechanisms will lead to more-realistic mechanisms in more-specific contexts.

6 Extensions

6.1 Partial Implementation

For partial implementation, the desideratum is a mechanism in which *some* equilibrium yields the socially optimal outcome. The corresponding privacy-protecting variant is the following:

Definition 6.1 (partial information-sensitive implementation at ψ) *A mechanism (H, A, g) is a* partial information-sensitive implementation *of an SCF f at ψ if there exists a strategy profile s^* for which the following hold:*

(a) s^* is an information-sensitive subgame perfect equilibrium at ψ, and

(b) $g(H(s^*(R^\psi))) = f(R)$ for all $R \in \mathcal{R}$.

Definition 6.2 (partial privacy-protecting implementation at ψ) *A mechanism (H, A, g) is a* partial privacy-protecting subgame perfect implementation *of an SCF f at ψ if it is a partial information-sensitive implementation, and if the guaranteed strategy s^* also satisfies the following:*

(c) For all $R, \overline{R} \in \mathcal{R}$ such that $f(R) = f(\overline{R})$, it holds that $H(s^*(\mathcal{R}^\psi)) = H(s^*(\overline{R}^\psi))$.

In the standard setup, the following is a trivial mechanism for partial implementation with complete information:

- Each agent i submits an outcome $a_i \in \mathcal{O}$.
- If at least $N - 1$ agents agree on an outcome a, then the outcome is a.
- Otherwise, the outcome is some arbitrary element of $\mathcal{O}$.

The strategy profile s^* is for every agent i to submit the element $a_i = f(R)$ at R. This is an equilibrium since no agent can alter the outcome. Note, however, that this may not be a partial information-sensitive implementation for lexicographic ψ. The problem is, what is the set of possible types following a unilateral deviation? For example, consider some R and a such that $f(R) = a$. Then $\mathrm{PT}^\psi(a, \ldots, a) = \mathcal{R}|_a$. But what about $\mathrm{PT}^\psi(b, a, \ldots, a)$ for some $b \neq a$? Even if ψ is one-deviation consistent, it is possible that $\mathrm{PT}^\psi(b, a, \ldots, a) = S \subsetneq \mathcal{R}|_a$. Furthermore, there is a lexicographic

ψ for which $(a, S)P_i^{\psi}(a, \mathcal{R}|_a)$. In this case, agent 1 would have a profitable unilateral deviation, and so the trivial mechanism may not be a partial implementation.

However, there is another mechanism that is a partial information-sensitive implementation and which is not much more complicated:

Proposition 6.3 *If $N \geq 3$, then there is a partial information-sensitive implementation of f for any ψ.*

The mechanism that achieves this implementation is the following:

- Each agent i submits a profile $R^i \in \mathcal{R}$.
- If at least $N - 1$ agents agree on a profile R, then the outcome is $f(R)$.
- Otherwise, the outcome is some arbitrary element of $\mathcal{O}$.

The strategy profile s^* is for every agent i to submit $R_i = R$ at R^{ψ}. This is an equilibrium since no agent can alter the outcome *nor* the set of possible types by a deviation. In particular, at any R^{ψ} the set PT^{ψ} resulting from the equilibrium strategy or from a unilateral deviation from it will always be $\{R\}$.

Observe that in the mechanism above, all private information is revealed. So what about partial privacy-protecting implementation, in which only the outcome $f(R)$ is revealed, and not all of R? Here I have the following impossibility result:

Theorem 6.4 *For any domain $\mathcal{R}$ with $|\mathcal{R}| \geq 3$ and any SCF $f : \mathcal{R} \mapsto \{0, 1\}$ there exists a lexicographic ψ that satisfies the following: there is no normal-form mechanism that is a privacy-protecting implementation of f at ψ.*

6.2 Information-Limiting Implementation

For privacy-protecting implementation the goal is to design a mechanism in which the only information revealed is the outcome, and not any additional private information beyond that. In this section we explore a more general notion in which agents may wish to reveal some or all information.

First, observe that it is impossible to reveal any *less* information than what is revealed in privacy-protecting implementation, while at the same time correctly implementing an SCF. This is simply because the correct outcome, together with the knowledge that it is the correct outcome, imply something about agents' preferences—namely, that the preferences are such that the SCF yields whatever outcome was

implemented. Next, observe that the most information that can be revealed is the entire set of preferences R. Thus, in this section we will consider implementation with revelation of information that is anywhere from being as fine as full revelation to as coarse as only revealing the outcome.

Now, it could be that in some profile of preferences R, an agent wants revelation of information, whereas in another profile $\overline{R}$ he desires privacy—i.e., the observer is unable to distinguish between R and $\overline{R}$. However, these are clearly impossible goals to achieve simultaneously (if in state R the observer learns the true state, then if he does not learn it he can deduce that the state is *not* R, violating the privacy desideratum). What can be achieved is the revelation and concealment of information according to a partition of $\mathcal{R}$. To that end, we will utilize the following notation: We will denote by Π a partition of $\mathcal{R}$ and by $\Pi(R) \subseteq \mathcal{R}$ the element of Π that includes R.

Definition 6.5 (information-limiting implementation) *A mechanism (H, A, g) is an* information-limiting implementation *of an SCF f with respect to a partition Π if it is an information-sensitive implementation of f, and if the strategy profile s^* guaranteed in Definition 2.6 also satisfies the following:*

1. *(c) For all $R \in \mathcal{R}$ it holds that if $\overline{R} \in \Pi(R)$, then $H(s^*(\mathcal{R}^\psi)) = H(s^*(\overline{R}^\psi))$, but if $\overline{R} \notin \Pi(R)$, then $H(s^*(\mathcal{R}^\psi)) \neq H(s^*(\overline{R}^\psi))$.*

Condition 1(c) states that the terminal history $H(s^*(R^\psi))$ could have been reached by s^* with any profile of preferences $\overline{R}^\psi$ for which $\overline{R} \in \Pi(R)$. Thus, the planner or any outside observer who sees the outcome $H(s^*(R^\psi)$ cannot differentiate between the true intrinsic preferences being R or $\overline{R}$. Unlike the case of privacy-protecting implementation, however, the planner can differentiate between R and $\overline{R}$ satisfying $f(R) = f(\overline{R})$ if $\overline{R} \notin \Pi(R)$.

Before stating the theorem we need a couple of definitions. The first is an appropriate variant of privacy favoring ψ.

Definition 6.6 (Π-favoring) *ψ is* Π-favoring *if for each $i \in N$ and $R \in \mathcal{R}$, it holds that*

$$(a, \Pi(R)) \; R_i^\psi \; (a, \{R\}).$$

The next definition requires that the partition be a refinement of the partition $\{\mathcal{R}|_{a,f}\}_{a \in \mathcal{O}}$. As discussed above, this is a necessary condition for implementation.

Definition 6.7 (f-consistent) *A partition* Π *is* f-consistent *if for each* $R \in \mathcal{R}$ *and* $\overline{R} \in \Pi(R)$*, it holds that* $f(R) = f(\overline{R})$*.*

The information-limiting variant of Theorem 5.2 is the following:

Theorem 6.8 *If there is a subgame perfect implementation of an SCF* $f : \mathcal{R} \mapsto \mathcal{O}$ *that satisfies NVP and* $N \geq 3$*, then there is an information-limiting implementation of* f *with respect to an* f*-consistent partition* Π *for any lexicographic,* Π*-favoring* ψ*.*

The information-limiting variant of Theorem 5.5 is the following:

Theorem 6.9 *If* $N \geq 3$ *and the SCF* $f : \mathcal{R} \mapsto \mathcal{O}$ *satisfies* $\alpha^{\max}$ *and NVP, then there is an information-limiting implementation of* f *with respect to an* f*-consistent partition* Π *for any MWR,* Π*-favoring* ψ*.*

6.3 When the Planner Does Not Know which Strategy Profile Is Played

One of the implicit assumptions in the definitions of information-sensitive and privacy-protecting implementation is that the planner knows the strategy profile being played. That is, the second requirement of these definitions is that for all profiles s that form an information-sensitive SPE, the outcome should be the same as dictated by f. However, the notion of an information-sensitive SPE relies on the sets PT, which of course depend on the profile s being played. Thus, a natural question is, what if the planner does not know the profile s? Perhaps he has a probabilistic belief about the profile s being played, or perhaps his uncertainty over which s is played is Knightian. In such situations, we may want a stronger notion of implementation. In particular, we may want the following requirement: for any profile s being played, if the outcome is not the same as dictated by f, then some player should have a profitable deviation from this profile *regardless of the beliefs of the planner about the profile being played.* This is formally captured by bullet 2 of the following definition.

Definition 6.10 (strong information-sensitive implementation) *A mechanism* (H, A, g) *is a* strong privacy-protecting subgame perfect implementation *of an SCF* f *at* ψ *if:*

1. *There exists a strategy profile* s^* *for which the following hold:*

   (a) s^* *is an information-sensitive subgame perfect equilibrium at* ψ*, and*

(b) $g(H(s^*(R^\psi))) = f(R)$ *for all* $R \in \mathcal{R}$.

(c) *For all* $R, \overline{R} \in \mathcal{R}$ *such that* $f(R) = f(\overline{R})$, *it holds that* $H(s^*(\mathcal{R}^\psi)) = H(s^*(\overline{R}^\psi))$.

2. *For all* $R \in \mathcal{R}$ *and strategy profiles* s *for which* $g(H(s(R))) \neq f(R)$, *there exists an agent* $i \in N$, *a history* $h \in H$, *and an* R^ψ*-deviation* s_i' *of agent* i *such that*

$$\left(g\left(H(s_{-i} \circ s_i'(R^\psi)|_h)\right), T\right) \; P_i^\psi \; \left(g\left(H(s(R^\psi)|_h)\right), S\right)$$

for all sets $S, T \subseteq \mathcal{R}$.

Note that the proofs of Propositions 5.1 and 5.4, as well as Theorems 5.2 and 5.5 go through with this stronger definition of implementation.

6.4 Non-singleton Ψ

In this paper we mainly considered the situation in which $R \in \mathcal{R}$ was unknown to the planner, but ψ was common knowledge. In this section I show how to extend our positive results to the case in which ψ is also unknown to the planner, and only the space of privacy types Ψ is known. The agents all know both R and ψ.

I first extend some of the definitions to handle a non-singleton Ψ. Let $\mathcal{R}^\Psi = \{R^\psi : R \in \mathcal{R}, \psi \in \Psi\}$. Recall that in a given mechanism (H, A, g), a pure strategy s_i of agent i in a mechanism (H, A, g) is a function that maps any pair (R^ψ, h) to an action from $A_i(h)$. Call a strategy s_i *privacy independent* if the strategy does not depend on ψ, which means that $s_i(R^\psi, h) = s_i(R^{\psi'}, h)$ for all $\psi, \psi' \in \Psi$ and all $h \in H$.

A privacy-independent profile s is an information-sensitive subgame perfect equilibrium at Ψ if it is an information-sensitive subgame perfect equilibrium at every $\psi \in \Psi$. The definition of a privacy-protecting implementation is also modified accordingly:

Definition 6.11 (privacy-protecting implementation at Ψ) *A mechanism* (H, A, g) *is a* privacy-protecting subgame perfect implementation *of an SCF* f *at* Ψ *if:*

1. *There exists a privacy-independent strategy profile* s^* *for which the following hold:*

   (a) s^* *is an information-sensitive subgame perfect equilibrium at every* $\psi \in \Psi$, *and*

(b) $g(H(s^*(R^\psi))) = f(R)$ *for all* $R \in \mathcal{R}$ *and* $\psi \in \Psi$.

(c) *For all* $\psi \in \Psi$ *and* $R, \overline{R} \in \mathcal{R}$ *such that* $f(R) = f(\overline{R})$, *it holds that* $H(s^*(R^\psi)) = H(s^*(\overline{R}^\psi))$.

2. *For all* $R \in \mathcal{R}$ *and strategy profiles* s *that form an information-sensitive subgame perfect equilibrium at any* $\psi \in \Psi$, *it holds that* $g(H(s(R^\psi))) = f(R)$.

The restrictions on information-sensitive preferences also directly translate to a setting with non-singleton Ψ. In particular, Ψ is lexicographic/MWR/privacy favoring if every $\psi \in \Psi$ is lexicographic/MWR/privacy favoring.

Again, note that the proofs of Propositions 5.1 and 5.4, as well as Theorems 5.2 and 5.5, go through with a non-singleton Ψ.

6.5 Social Choice Correspondences

In this section I extend the definition of implementation and the negative result of Theorem 4.2 to SCCs. We first need a definition.

Definition 6.12 (restrictions of an SCC) *A* restriction *of an SCC* $F : \mathcal{R} \twoheadrightarrow \mathcal{O}$ *is any function from the set* $\{f : \mathcal{R} \mapsto \mathcal{O}$ *such that* $f(R) \in F(R)\ \forall R \in \mathcal{R}\}$.

The following definition is very similar to Definition 2.7.

Definition 6.13 (privacy-protecting implementation of an SCC) *A mechanism* (H, A, g) *is a* privacy-protecting subgame perfect implementation *of an SCC* F *at* ψ *if:*

1. *There exists a restriction* f *of* F *and a strategy profile* s^* *for which the following hold:*

   (a) s^* *is an information-sensitive subgame perfect equilibrium at* ψ, *and*

   (b) $g(H(s^*(R^\psi))) = f(R)$ *for all* $R \in \mathcal{R}$.

   (c) *For all* $R, \overline{R} \in \mathcal{R}$ *such that* $f(R) = f(\overline{R})$, *it holds that* $H(s^*(\mathcal{R}^\psi)) = H(s^*(\overline{R}^\psi))$.

2. *For all* $R \in \mathcal{R}$ *and strategy profiles* s *that form an information-sensitive subgame perfect equilibrium at* ψ, *it holds that* $g(H(s(R^\psi))) \in F(R)$.

This definition is slightly different from the standard definition of implementation of SCCs. The standard definition, such as the one described in Section 2.5, requires that for all R, *any* outcome in $F(R)$ be obtainable in equilibrium. In order to make this compatible with the information-sensitive or privacy-protecting versions of implementation, we could have required that bullet 1 in Definition 6.13 hold for all restrictions f of F, and not just for one particular restriction. Such a stronger definition, however, would have a few subtleties that would make it difficult to work with. Additionally, since I will only be generalizing our negative result to the case of SCCs, a weaker definition makes the negative result stronger.

Before extending Theorem 4.2 to SCCs, we need the following definition. An SCC F is a *dictatorship* if there exists an agent i such that $F(R)$ is always equal to the set of elements ranked highest by agent i, i.e., $F(R) \equiv \mathrm{TR}_i(R)$.

Theorem 6.14 *For any pairwise-rich $\mathcal{R}$, if there exists a normal-form mechanism that is a privacy-protecting Nash implementation of an SCC $F : \Theta \twoheadrightarrow \{0,1\}$ at $\mathcal{R}^0$ then F is either constant or a dictatorship.*

Appendix

A Proof of Proposition 3.1

Proof of Proposition 3.1: Fix an SCF $f : \mathcal{R} \mapsto \mathcal{O}$, and suppose towards a contradiction that the mechanism (H, A, g) is an information-sensitive subgame perfect implementation of some non-constant f at ψ. Since f is non-constant, there exist distinct $a, b \in \mathcal{O}$ and intrinsic preferences $R^a, R^b \in \mathcal{R}$ such that $f(R^a) = a$ and $f(R^b) = b$. Let s^* be the strategy profile guaranteed by Definition 2.6.

Consider now the strategy profile s, such that $s(R) \equiv s^*(R^a)$ for all $R \in \mathcal{R}$. Also, for every nonterminal history $h \in H$, let $\mathrm{PT}_h(H(s^*(R^a)|_h), s) = \mathcal{R}$. In words, these sets of possible types mean that starting at any history, if a terminal history is reached that could have been reached by s, then all intrinsic preferences are possible. This is reasonable, since s is the same regardless of the intrinsic preferences. We will define $\mathrm{PT}_h(z, s)$ for other z's in the sequel.

We claim that the profile s constitutes an information-sensitive subgame perfect equilibrium at ψ (when the sets PT are as will be defined). This, of course, implies a contradiction, since s does *not* always yield the social optimum according to f

in (H, A, g). In particular, s always yields the outcome a, but there are intrinsic preferences in which the socially optimal outcome is b (namely, R^b).

We now show that s is an information-sensitive subgame perfect equilibrium at ψ. Suppose towards a contradiction that this is not the case, and that there is some nonterminal history h, some $i \in N$, some $R \in \mathcal{R}$, and some R^ψ-deviation s_i' of agent i for which

$$(g(z'), \mathrm{PT}_h(z', s)) \;\; P_i^\psi \;\; (g(z), \mathrm{PT}_h(z, s)) \,,$$

where $z = H(s(R^\psi)|_h)$ and $z' = H(s_{-i} \circ s_i'(R^\psi)|_h)$.

Recall our assumption above that $\mathrm{PT}_h(z, s) = \mathcal{R}$, yielding

$$(g(z'), \mathrm{PT}_h(z', s)) \;\; P_i^\psi \;\; (g(z), \mathcal{R}) \,.$$

The crucial question now is, what should $\mathrm{PT}_h(z', s)$ be?

Observe that regardless of $\mathrm{PT}_h(z', s)$, it must be the case that $g(z') P_i^\psi g(z)$. For otherwise, if $g(z') R_i^\psi g(z)$, then it follows that

$$(g(z'), \mathrm{PT}_h(z', s)) \;\; R_i^\psi \;\; (g(z), \mathcal{R}) \,,$$

since the only way to get a strict improvement over an outcome and no information revelation is to get a better outcome and no information revelation. However, now a planner can "learn" that agent i will deviate *only* if this yields him a strictly better outcome. In other words, the planner now learns that the true intrinsic preferences must lie in the set $\mathcal{R}' = \{R \in \mathcal{R} : g(z') P_i g(z)\} \subsetneq \mathcal{R}$! Thus, defining the set $\mathrm{PT}_h(z', s) \stackrel{\text{def}}{=} \mathcal{R}'$ yields the contradiction that

$$(g(z'), \mathcal{R}') \;\; P_i^\psi \;\; (g(z), \mathcal{R}) \,.$$

Hence, s is an information-sensitive subgame perfect equilibrium. ■

B Proofs from Section 4

Proof of Theorem 4.1: This theorem follows directly from Theorem 6.4. ■

Next, we restate the definition of outcome monotonicity for the case of SCCs. For the relevant definitions related to SCCs see Section 6.5.

Definition B.1 (outcome monotonicity (OM)) *an SCC F satisfies* outcome monotonicity (OM) *if there is some restriction f of F such that for any $\overline{R} \in \mathcal{R}$ and outcome $a \notin F(\overline{R})$, there exists an agent $i \in N$ and an outcome $b \in \mathcal{O}$ for which $b \overline{P}_i a$ but $a R_i b$ for all $R \in \mathcal{R}$ satisfying $a = f(R)$.*

We now restate Lemma 4.5 for SCCs.

Lemma B.2 *If there exists a normal-form mechanism that is a privacy-protecting Nash implementation of an SCC F at $\mathcal{R}^o$, then F satisfies OM.*

Proof: Let f be as in Definition 6.13, or, if F is a function, let $f \equiv F$. Observe that in a normal-form mechanism, condition 1(c) of Definitions 2.7 and 6.13 is equivalent to the requirement that s^* depend only on $f(R)$, and not on all of R. That is, for any $R, \overline{R} \in \mathcal{R}$ with $f(R) = f(\overline{R})$ it should be the case that $s^*(R^o) = s^*(\overline{R}^o)$.

Fix some $R \in \mathcal{R}$ and $a \notin F(R)$. Also fix some $\overline{R} \in \mathcal{R}$ for which $a = f(\overline{R})$. Let s^* be the information-sensitive Nash equilibrium guaranteed in a privacy-protecting Nash implementation (H, A, g) of F. We must have $g(s^*(R^o)) = a$ and $g(s^*(\overline{R}^o)) \neq a$.

Consider now the profile of strategies s that is identical to s^* at every state except at R^o, and set $s(R^o) = s^*(\overline{R}^o))$. Observe that s cannot be an information-sensitive Nash equilibrium, since $g(s(R^o)) = g(s^*(\overline{R}^o)) = a$ even though $a \notin F(R)$ (and so this contradicts condition 2 of Definition 6.13). Thus, there exist $i \in N$, $R \in \mathcal{R}$, and a strategy s_i' that yields agent i a higher payoff than s_i. However, since s differs from s^* only at R^o, and agent i does not care about privacy (since his information-sensitive preferences are R_i^o), it must be the case that i's improved payoff is obtained at R^o. That is, fixing $g(s_{-i} \circ s_i'(R^o)) = b$ we have

$$(b, \mathrm{PT}(b, s)) \; P_i^o \; (a, \mathrm{PT}(a, s)) \,,$$

which implies that

$$bP_i a \tag{1}$$

again since i's information-sensitive preferences are R_i^o.

Now consider some $\overline{R} \in \mathcal{R}$ satisfying $a = f(\overline{R})$. By our observation at the beginning of the proof, it must be the case that $s^*(\widetilde{R}^o) = s^*(\overline{R}^o)$, since $f(\widetilde{R}) = f(\overline{R}) = a$. Consider a deviation s_i'' of agent i, where s_i'' is identical to i's strategy in s^* at every state except $(\widetilde{R}^o)$, and where $s_i''(\widetilde{R}^o) = s_i'(\overline{R}^o)$. The strategy profile $s_{-i}^* \circ s_i''$ yields the same outcomes in every state as s^* except in state $(\widetilde{R}^o)$, where it yields outcome b (since $g(s_{-i} \circ s_i'(\overline{R}^o)) = b$).

However, since s^* *is* a privacy-protecting Nash equilibrium, the deviation s_i'' cannot be beneficial to agent i. That is,

$$(a, \mathrm{PT}(a, s^*)) \; \widetilde{R}_i^o \; (b, \mathrm{PT}(b, s^*)) \,,$$

which implies that

$$a\widetilde{R}_i^o b \tag{2}$$

since i's information-sensitive preferences are R_i^o.

Finally, equation (1), together with the fact that (2) holds for every $\widetilde{R}$ with $a = f(\widetilde{R})$, implies that the SCC F satisfies OM. ■

We now restate Lemma 4.6 for SCCs.

Lemma B.3 *For any pairwise-rich $\mathcal{R}$, a non-constant SCC $F : \mathcal{R} \twoheadrightarrow \{0,1\}$ satisfies outcome monotonicity if and only if it is a dictatorship.*

Proof: Let f be the restriction of F guaranteed by Definition B.1. Fix any $\overline{R} \in \mathcal{R}$ for which $0 \notin F(\overline{R})$, and denote $\mathcal{R}_0 = \{R \in \mathcal{R} : f(R) = 0\}$. Note that, since F is non-constant, such a $\overline{R}$ must exist and $\mathcal{R}_0$ must be nonempty. Outcome monotonicity implies that there exists an agent $i \in N$ for which $1\overline{P}_i 0$ but $0R_i 1$ for all $R \in \mathcal{R}_0$. In particular, for all $R \in \mathcal{R}$ satisfying $1P_i 0$ it must be the case that $f(R) = 1$: Otherwise we would have $R \in \mathcal{R}_0$, which would imply the contradiction that $0R_i 1$.

Now fix any $\widetilde{R} \in \mathcal{R}$ for which $1 \notin F(\widetilde{R})$ and denote $\mathcal{R}_1 = \{R \in \mathcal{R} : f(R) = 1\}$. Again, such a θ_1 must exist and $\mathcal{R}_1$ must be nonempty since F is non-constant. Outcome monotonicity implies that there exists an agent $j \in N$ for which $0\widetilde{P}_j 1$ but $1R_j 0$ for all $R \in \mathcal{R}_1$. In particular, for all $R \in \mathcal{R}$ satisfying $0P_j 1$ it must be the case that $f(R) = 0$: Otherwise we would have $R \in \mathcal{R}_1$, which would imply the contradiction that $1R_j 0$.

Suppose that $i \neq j$, and consider the preference profile R^{ij} for which $1P_i^{ij}0$ and $0P_j^{ij}1$. Such a R^{ij} must exist in $\mathcal{R}$ since the latter is pairwise-rich. The above implies that we must have both $f(R^{ij}) = 1$ and $f(R^{ij}) = 0$, a contradiction. Thus, $i = j$.

Now consider a profile $R^{\text{ind}} \in \mathcal{R}$ for which $0I_i^{\text{ind}}1$. If there does not exist such a profile or $F(R^{\text{ind}}) = \{0,1\}$ for all such profiles, then F is a dictatorship, and the lemma is proved.

Otherwise, suppose that $F(R^{\text{ind}}) = \{1\}$ (the case of $F(R^{\text{ind}}) = \{0\}$ is symmetric, but otherwise identical). Again, outcome monotonicity implies that there exists an agent $k \in N$ for which $1P_k^{\text{ind}}0$ but $0R_k 1$ for all $R \in \mathcal{R}_0$. In particular, for all $R \in \mathcal{R}$ satisfying $1P_k 0$ it must be the case that $f(R) = 1$: Otherwise we would have $R \in \mathcal{R}_0$, which would imply the contradiction that $0R_k 1$.

Note that k cannot be equal to i, since $0I_i^{\text{ind}}1$ but $1P_k^{\text{ind}}0$. Consider then the profile $R^{ik} \in \mathcal{R}$ for which $1P_k^{ik}0$ and $0P_i^{ik}1$. Again such a profile must exist in $\mathcal{R}$ since the

domain is pairwise-rich. The above implies that we must have both $f(R^{ik}) = 1$ and $f(R^{ik}) = 0$, a contradiction. Thus, if $0 I_i^{\text{ind}} 1$ then $F(R^{\text{ind}}) = \{0, 1\}$, and so F must be a dictatorship.

For the reverse implication, it is clear that the dictatorship function satisfies outcome monotonicity. Any restriction of F works. ■

C Proofs from Section 5

Our possibility results about implementation with extensive-form mechanisms are constructive, and the mechanism we use are variants of the mechanism of Moore and Repullo (1988).

The Moore-Repullo mechanism that implements an SCC F in subgame perfect equilibrium, and which is also used by Abreu and Sen (1990), uses sequences of agents $j(0), \ldots, j(\ell)$ and a sequences of outcomes $a_0, \ldots, a_{\ell+1}$, one such pair of sequences for each $R \in \mathcal{R}$, $\overline{R} \in \mathcal{R}$, and $a \in F(R) - F(\overline{R})$, satisfying Condition α (see Definition 2.4). The mechanism is the following (copied almost verbatim from Abreu and Sen (1990)):

The Moore-Repullo Mechanism (MR):

- **Stage 0:** Each agent i simultaneously submits a triplet $(R^i, a^i, n^i) \in \mathcal{R} \times \mathcal{O} \times \mathbb{Z}$. If $N-1$ agents submit the same R and $a \in F(R)$ then the outcome is a, unless the non-agreeing agent j announces R^j with $a \in F(R) - F(R^j)$ and $j = j(0)$ in the sequence $j(R, R^j, a)$. In this latter case, go to Stage 1.

  In all other cases the agent who announced the highest integer selects any outcome in $\mathcal{O}$.

- **Stage k, $k = 1, \ldots, \ell$:** Each agent i simultaneously either raises a "flag" or announces a nonnegative integer.

  If at least $N-1$ agents raise flags, the agent $j(k-1)$ (in the sequence $j(R, R^j, a)$) selects any outcome in $\mathcal{O}$.

  If at least $N-1$ agents announce 0 the outcome is a_k, unless $j(k)$ does not announce 0, in which case go to the next stage, or, if $k = \ell$, implement $a_{\ell+1}$.

  In all other cases the agent who announced the highest integer selects any outcome in $\mathcal{O}$.

Abreu and Sen (1990) show that the following strategy profile s^{MR} is a SPE of MR: In Stage 0, $s^{MR}(R, \epsilon) = (R, a, 0)$, where $a \in F(R)$. In all subsequent stages all agents always announce 0. They also prove the following as part of the proof of their Theorem 2.

Lemma C.1 *Suppose the true strategy profile is R, but all agents submit $R^i = \overline{R} \neq R$ and outcome $a \notin F(R)$ in Stage 0 of the Moore-Repullo mechanism. Then if agent $j(0)$ of the sequence $j(\overline{R}, R, a)$ deviates and submits $R^i = R$, then all SPE outcomes following this deviation are in $TR_{j(0)}(R)$.*

Consider now the following mechanism MR$'$, which is a slight variation of the Moore-Repullo mechanism:

Mechanism MR$'$:

- **Stage 0:** Same as the Moore-Repullo mechanism.
- **Stage k, $k = 1, \ldots, \ell$:** Same as the Moore-Repullo mechanism, except that in each stage k, each agent i also submits a vector of preferences $R^i \in \mathcal{R}$.

The MR$'$ mechanism is almost identical to the Moore-Repullo mechanism, except for the expanded action space in stages 1 through ℓ. Note that these additions do not impact the outcome of the mechanism. Consider the strategy profile $s^{\text{MR}'}$ that is identical to s^{MR}, except that in all stages following 0 the agents submit the true profile of preferences R (in addition to announcing 0). Observe that $s^{\text{MR}'}$ is an SPE of the mechanism MR$'$. We use mechanism MR$'$, together with the strategy profile $s^{\text{MR}'}$, in the proof of Proposition 5.1. But first, we need some definitions and lemmas.

C.1 Some Technical Definitions and Lemmas

Definition C.2 (maximally-dispersed) *A strategy profile s in a mechanism (H, A, g) is* maximally-dispersed *at ψ if for every $i \in N$, $R \in \mathcal{R}$, $h \in H \setminus Z$, and R^ψ-deviation $s_i' \neq s_i$ of agent i it holds that*

(i) $\text{LD}_h\left(H\left(s_{-i} \circ s_i'\left(R^\psi\right)|_h\right), s\right) = \{R\}$ *and*

(ii) $\text{LD}_h\left(H\left(s\left(R^\psi\right)|_h\right), s\right) = \{R\}$.

Definition C.3 (outcome-condensed) *A strategy profile s in a mechanism (H, A, g) is* outcome-condensed *at ψ if for every $i \in N$, $R \in \mathcal{R}$, $h \in H \setminus Z$, and R^ψ-deviation $s_i' \neq s_i$ of agent i it holds that*

(i) $\mathrm{LD}_h\left(H\left(s_{-i} \circ s_i'\left(R^\psi\right)|_h\right), s\right) = \{R\}$ *and*

(ii)' $g(s(R^\psi)|_h) = g(s_{-i} \circ s_i'(R^\psi)|_h) = a$ *for some* $a \in \mathcal{O}$ *and*

$$\mathrm{PT}_h\left(H\left(s\left(R^\psi\right)|_h\right), s\right) = \mathcal{R}|_a.$$

Lemma C.4 *Let s^o be a subgame perfect equilibrium of a mechanism (H, A, g), and let s be the strategy profile satisfying $s_i(R^\psi, h) = s_i^o(R, h)$ for some ψ and all i, R, and h. Then if either s is outcome-condensed and ψ is privacy favoring or s is maximally-dispersed then s is also an information-sensitive subgame perfect equilibrium at ψ.*

Proof: Fix an agent i, a profile of intrinsic preferences $R \in \mathcal{R}$, and some $h \in H \setminus Z$. We will show that agent i does not have a unilateral R^ψ-deviation from s at h that will yield him a strictly higher payoff.

Let s_i' be some R^ψ-deviation of agent i from s, and suppose towards a contradiction that

$$(g(z'), \mathrm{PT}_h(z', s)) \; P_i^\psi \; (g(z), \mathrm{PT}_h(z, s)), \tag{3}$$

where $z' = H(s_{-i} \circ s_i'(R^\psi)|_h)$ and $z = H(s(R^\psi)|_h)$. Since $\mathrm{PT}_h(z', s) \subseteq \mathrm{LD}_h(z', s)$ by the second restriction, and since s is maximally-dispersed or outcome-condensed it follows that $\mathrm{PT}_h(z', s) = \{R\}$.

Now, since s^o is a subgame perfect equilibrium (SPE), it cannot be the case that $z' P_i z$. If it were, then s_i' would be a profitable local deviation from s^o at R, which contradicts the assumption that s^o is a SPE.

Thus, it must be the case that $z R_i z'$. If s is maximally-dispersed, and so $\mathrm{PT}_h(z, s) = \{R\}$, then $z R_i z'$ implies that

$$(g(z), \{R\}) \; R_i^\psi \; (g(z'), \{R\}).$$

Thus,

$$(g(z), \mathrm{PT}_h(z, s)) \; R_i^\psi \; (g(z'), \mathrm{PT}_h(z', s)),$$

contradicting (3) above.

Alternatively, if s is maximally-condensed, and so $\mathrm{PT}_h(z,s) = \mathcal{R}|_a$ and $z = z' = a$, then since ψ is privacy favoring it must be the case that

$$(a, \mathcal{R}|_a) \; R_i^\psi \; (a, \{R\}).$$

Thus, once again

$$(g(z), \mathrm{PT}_h(z,s)) \; R_i^\psi \; (g(z'), \mathrm{PT}_h(z',s)),$$

contradicting (3) above.

Thus, there can be no beneficial R^ψ-deviation of agent i at h, and so s is an information-sensitive subgame perfect equilibrium of (H, A, g) at ψ. ■

Lemma C.5 *In a mechanism (H, A, g), if s is a information-sensitive subgame perfect equilibrium at some lexicographic ψ, then the profile s^o satisfying $s_i^o(R,h) = s_i(R^\psi, h)$ for all i, R, and h is a subgame perfect equilibrium of (H, A, g).*

Proof: Suppose towards a contradiction s^o is not a SPE. This implies that there exists an agent $i \in N$, a history $h \in H \setminus Z$, a profile $R \in \mathcal{R}$, and an R-deviation s_i' of agent i such that $g(z')P_i g(z)$, where $z' = H(s_{-i}^o \circ s_i'(R))|_h$ and $z = H(s^o(R))|_h$.

Let s_i'' be the R^ψ-deviation of agent i from s that satisfies $s_i''(R^\psi, h) = s_i'(R,h)$ for all R and h, and observe that

$$z'' = H(s_{-i} \circ s_i''(R^\psi)|_h)$$

and

$$z = H(s(R^\psi)|_h).$$

Thus, $g(z'')P_i^\psi g(z)$. Furthermore, since ψ is lexicographic it also holds that

$$(g(z''), S) \; P_i^\psi \; (g(z), T)$$

for any sets S and T with $R \in T$. The first requirement on the sets PT implies that $R \in \mathrm{PT}_h(z,s)$. Thus, since $T = \mathrm{PT}_h(z,s)$, we get that

$$(g(z''), \mathrm{PT}_h(z'',s)) \; P_i^\psi \; (g(z), \mathrm{PT}_h(z,s)),$$

contradicting the assumption that s is an information-sensitive subgame perfect equilibrium at ψ. Hence, s^o is a SPE of (H, A, g). ■

We also have a similar lemma for the case of MWR preferences, but first need a definition.

Definition C.6 (maximal deviability) *A mechanism (H, A, g) satisfies* maximal deviability *with respect to an SCC F if the following holds for any $R \in \mathcal{R}$, profile s satisfying $g(H(s(R))) \notin F(R)$, and history h: Under R, either s is a Nash equilibrium of the subgame rooted at h, or some agent i has a deviation s_i' that will yield an outcome in $\mathrm{TR}_i(R)$.*

Lemma C.7 *Fix a mechanism (H, A, g) that satisfies maximal deviability with respect to an SCC F, and suppose the profile s is a information-sensitive subgame perfect equilibrium at some MWR ψ. Then for every $R \in \mathcal{R}$, either $g(H(s(R^\psi))) \in F(R)$ or the profile $s^o(R)$ satisfying $s_i^o(R, h) = s_i(R^\psi, h)$ for all i and h is a subgame perfect equilibrium of (H, A, g) at R.*

Proof: Suppose towards a contradiction that for some $R \in \mathcal{R}$ it holds that both $g(H(s(R^\psi))) \notin F(R)$ and that $s^o(R)$ is not a SPE at R. Observe that $g(H(s(R^\psi))) = g(H(s^o(R)))$, and so $g(H(s^o(R))) \notin F(R)$. This implies that there exists an agent $i \in N$, a history $h \in H \setminus Z$, and an R-deviation s_i' of agent i such that $z' P_i z$, where $z' = g(H(s^o_{-i} \circ s_i'(R))|_h)$ and $z = g(H(s^o(R))|_h)$. Since (H, A, g) satisfies maximal deviability with respect to F and $g(H(s^o(R))) \notin F(R)$, we can choose i, h, and s_i' in such a way that $z' \in \mathrm{TR}_i(R)$.

Let s_i'' be the R^ψ-deviation of agent i from s that satisfies $s_i''(R^\psi, h) = s_i'(R, h)$ for all R and h, and observe that

$$z'' = H(s_{-i} \circ s_i''(R^\psi)|_h)$$

and

$$z = H(s(R^\psi)|_h).$$

Thus, $g(z'') P_i^\psi g(z)$ and $g(z'') \in \mathrm{TR}_i(R)$. Furthermore, since ψ satisfies MWR it also holds that

$$(g(z''), S)\ P_i^\psi\ (g(z), T)$$

for any sets S and T with $R \in T$. The first requirement on the sets PT implies that $R \in \mathrm{PT}_h(z, s)$. Thus, since $T = \mathrm{PT}_h(z, s)$, we get that

$$(g(z''), \mathrm{PT}_h(z'', s))\ P_i^\psi\ (g(z), \mathrm{PT}_h(z, s)),$$

contradicting the assumption that s is an information-sensitive subgame perfect equilibrium at ψ. Hence, s^o is a SPE of (H, A, g). ■

Lemma C.8 *Fix some ψ and suppose the number of agents $N \geq 3$. Then the strategy profile s for which $s(R^\psi, h) = s^{\mathrm{MR}'}(R, h)$ for all R and h is maximally-dispersed at ψ in the* MR' *mechanism.*

Proof: Fix some $i \in N$, $R \in \mathcal{R}$, and $h \in H \setminus Z$, and let $z = H(s\left(R^\psi\right)|_h)$. Observe that, since the number of agents is at least 3, and since all agents submit the same actions at every history, it is the case that $\mathrm{L}_h(z, s) = \mathrm{LD}_h(z, s)$. Furthermore, since there is exactly one $\overline{R} \in \mathcal{R}$ such that $H\left(s\left(\overline{R}^\psi\right)|_h\right) = z$ (namely, $\overline{R} = R$), it holds that $\mathrm{LD}_h(z, s) = \mathrm{L}_h(z, s) = \{R\}$.

Next, observe that when a single agent deviates from s at (R, h) to yield a terminal history z', it is always possible to uniquely determine R from h, s, and z'. This follows from the facts that $\mathrm{L}_h(z, s) = \{R\}$, that the number of agents is at least three, and that agents always submit the same actions. Thus, for any i and R^ψ-deviation $s_i' \neq s_i$ of agent i it holds that

$$\mathrm{LD}_h\left(H\left(s_{-i} \circ s_i'\left(R^\psi\right)|_h\right), s\right) = \{R\},$$

and so s is maximally-dispersed. ■

C.2 Proofs of Propositions 5.1 and 5.4

Proof of Proposition 5.1: By Theorem 1 of Abreu and Sen (1990), if an SCF f is subgame perfect implementable then f satisfies Condition α. Furthermore, by Theorem 2 of Abreu and Sen (1990), since f also satisfies NVP, it is implementable via the Moore-Repullo mechanism MR. It is then also implementable by mechanism MR', since the enlarged set of actions there has no effect on the outcomes of the mechanism.

Now, consider the strategy profile s^* for which $s^*(R^\psi, h) = s^{\mathrm{MR}'}(R, h)$ for all R and h. By Lemma C.8, this strategy profile is maximally-dispersed at ψ. By Lemma C.4, s^* is an information-sensitive subgame perfect equilibrium of MR'. In addition, by the properties of this strategy profile, $s^*(R^\psi) = f(R)$ for all $R \in \mathcal{R}$. Thus, bullets 1(a) and 1(b) of Definition 2.6 are satisfied.

Furthermore, for any information-sensitive subgame perfect equilibrium s in MR' at ψ, Lemma C.5 implies that there is a corresponding SPE with respect to o (since ψ is lexicographic by assumption). However, since MR' is a subgame perfect implementation of f, it must then be the case that $g(H(s(R))) = f(R)$ for all $R \in \mathcal{R}$. Since

$H(s^*(R^\psi)) = H(s(R))$ it holds that $g(H(s(R^\psi))) = f(R)$ for all $R \in \mathcal{R}$, satisfying bullet 2 of Definition 2.6.

Thus, MR′ is an information-sensitive implementation of the SCF f with respect to any lexicographic ψ. ■

Proof of Proposition 5.4: The mechanism we will use is MR′. Consider the strategy profile s^* for which $s^*(R^\psi, h) = s^{\mathrm{MR}'}(R, h)$ for all R and h. By Lemma C.8, this strategy profile is maximally-dispersed at ψ. By Lemma C.4, s is an information-sensitive subgame perfect equilibrium of MR′. In addition, by the properties of this strategy profile, $s^*(R^\psi) = f(R)$ for all $R \in \mathcal{R}$. Thus, bullets 1(a) and 1(b) of Definition 2.6 are satisfied.

Furthermore, Lemma C.9 below states that for any SCC F satisfying Condition $\alpha^{\max}$ and NVP, the Moore-Repullo mechanism, and hence also the mechanism MR′, satisfy maximal deviability with respect to F. So for any information-sensitive subgame perfect equilibrium s in MR′ with respect to ψ and any $R \in \mathcal{R}$, Lemma C.7 implies that either $g(H(s(R^\psi))) = f(R)$, or there is a corresponding SPE s^o at R with respect to o. However, since MR′ is a subgame perfect implementation of f, it must be the case that $g(H(s^o(R))) = f(R)$, and so also $g(H(s^o(R))) = f(R)$. Since $H(s^*(R^\psi)) = H(s(R))$ it holds that $g(H(s(R^\psi))) = f(R)$ for all $R \in \mathcal{R}$, satisfying bullet 2 of Definition 2.6.

Thus, MR′ is an information-sensitive implementation of the SCF f with respect to any MWR ψ. ■

Lemma C.9 *Under NVP, the implementation of an SCC F satisfying Condition $\alpha^{\max}$ using the Moore-Repullo mechanism satisfies maximal deviability with respect to F.*

Proof: The proof follows Theorem 2 of Abreu and Sen (1990), with the observation that whenever an agent has a deviation from some strategy profile s that does not lead to an outcome in F, he actually has a deviation that will yield an outcome that is top-ranked by him. ■

C.3 Proofs of Theorems 5.2 and 5.5

The proofs of Theorems 5.2 and 5.5 both use the following mechanism MRP:

Mechanism MRP:

- **Stage 0(a):** Each agent i simultaneously submits a pair $(a_i, n_i) \in \mathcal{O} \times \mathbb{Z}$. If $a_1 = \ldots = a_N$ then the outcome is $g(a_1, \ldots, a_N) = a_1$. If there exists $j_d \in N$ and $a \in \mathcal{O}$ such that $a_i = a$ for all agents $i \neq j_d$ but $a_{j_d} \neq a$, then go to Stage 0(b). Otherwise, the agent who submitted the highest integer n_i chooses the outcome of the mechanism.

- **Stage 0(b):** Suppose the history is $h = ((a_1, n_1), \ldots, (a_N, n_N))$. Each agent i simultaneously submits a pair $(R^i, n^i) \in \{\mathcal{R} \cup \{\perp\}\} \times \mathbb{Z}$.

  (i) If there does *not* exist an agent j and a profile $R \in \mathcal{R}$ such that $R^i = R$ for all $i \in N \setminus \{j\}$, then the agent who submitted the highest integer n^i chooses the outcome of the mechanism.

  (ii) If there *does* exist an agent j and a profile $R \in \mathcal{R}$ such that $R^i = R$ for all $i \in N \setminus \{j\}$ and if $a_i = f(R)$ for all $i \in N \setminus \{j, j_d\}$ then the outcome is $f(R)$, unless agent j announces R^j with $f(R^j) \neq f(R)$ and $j = j(0)$ in the sequence $j(R, R^j, f(R))$. In this latter case:

  1. Fix $a^i \stackrel{\text{def}}{=} a_i$ for all $i \in N \setminus \{j_d\}$;
  2. If $j \neq j_d$ then fix $a^{j_d} \stackrel{\text{def}}{=} a$; If $j = j_d$ then fix $a^{j_d} \stackrel{\text{def}}{=} a_{j_d}$.
  3. Go to stage 1.

  (iii) In all other cases, agent j_d chooses the outcome of the mechanism.

- **Stage k, $k = 1, \ldots, \ell$:** Continue with Stage k of the MR$'$ mechanism, where the Stage 0 history is $((R^1, a^1, n^1), \ldots, (R^N, a^N, n^N))$.

Consider the following strategy s^{MRP}:

- In Stage 0(a), $s_i^{\text{MRP}}(R, \epsilon) = (f(R), 0)$ for every $i \in N$.

- In Stage 0(b), profile R, and history h, every agent i submits $s_i^{\text{MRP}}(R, h) = (R, 0)$.

- For all subsequent rounds (of the MR$'$ part of the mechanism), agents always announce $(0, R)$, as in $s^{\text{MR}'}$.

We now proceed with various lemmas that will be used in the proofs of Theorems 5.2 and 5.5.

Lemma C.10 *If f satisfies α, then for every R the profile s^{MRP} is a subgame perfect equilibrium of* MRP *at R.*

Proof: First observe that s^{MRP} is an SPE in stages 1 and onwards: At history $((a_1, n_1), \ldots, (a_N, n_N), (R^1, n^1), \ldots, (R^N, n^N))$ and onwards, the strategy and mechanism are identical to the SPE strategy $s^{\text{MR}'}$ in MR$'$ at history $((R^1, a^1, n^1), \ldots, (R^N, a^N, n^N))$ and onwards.

Now consider stage 0(b) and a history $h = ((a_1, n_1), \ldots, (a_N, n_N))$. If there exists $j_d \in N$ such that $a_i = f(R)$ for all agents $i \neq j_d$, then the agents are essentially playing stage 0 of MR$'$ (this is case (i)). Thus, since s^o is identical to $s^{\text{MR}'}$ here, there is no profitable deviation. On the other hand, if there does not exist $j_d \in N$ such that $a_i = f(R)$ for all agents $i \neq j_d$, then, regardless of any unilateral deviation, s^{MRP} will lead to case (ii), and so the outcome will be chosen by agent j_d. To see this, observe that the history must be such that there exists some agent j_d and an outcome $b \neq f(R)$ such that $a_i = b$ for all $i \neq j_d$. If no agent deviates, then the outcome will be chosen by agent j_d, since it is not the case that $a_i = f(R)$ for any i, and so this leads to case (ii). If some agent does deviate, then the mechanism can still determine the true R from agents' messages (since $N - 1$ agents agree on the same R). However, even then it will not be the case that $a_i = f(R)$ for any agent i except possibly j_d. This, a unilateral deviation also leads to case (ii).

Finally, consider stage 0(a) of the mechanism. Observe that the strategy profile s^{MRP} here is identical to $s^{\text{MR}'}$. Thus, any profitable unilateral deviation here would imply a profitable deviation from $s^{\text{MR}'}$ in MR$'$. Since $s^{\text{MR}'}$ is an SPE profile, there is no profitable deviation from s^{MRP} in stage 1 either. ■

Lemma C.11 *If f satisfies α and NVP, then for any $R \in \mathcal{R}$ all subgame perfect equilibria of* MRP *at R lead to the outcome $f(R)$.*

Proof: Suppose towards a contradiction that there is a SPE s of (H, A, g) that, at R, leads to an outcome $b \neq f(R)$. Observe first that in stage 0(a) of the mechanism it must be the case that $s_i(R, \epsilon) = b$ for all $i \in N$. Otherwise, if not all agents agree on the same outcome b, then at least $N - 1$ agents have a unilateral deviation that would yield them a top-ranked outcome. Thus, b must be top-ranked by all these

agents. But if b is a top-ranked outcome for $N-1$ agents, then by NVP it must be the case that $b = f(R)$, a contradiction.

Now consider stage 0(b) with a history $h = ((a_1, n_1), \ldots, (a_N, n_N))$. Suppose that for every $i \in N$ the strategy is $s(R, h) = (R^i, n^i)$. There are two cases to consider:

1. There exists an $\overline{R} \in \mathcal{R}$ such that $R^i = \overline{R}$ for all i, and $b = f(\overline{R})$. In this case, we claim that agent $j(0)$ in the sequence $j(\overline{R}, R, b)$ has a profitable deviation that yields him a top-ranked outcome. In particular, he can deviate to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. This situation is then the same as the subgame of MR$'$ that follows the history h' in which all agents but $j(0)$ announce $(\overline{R}, b)$ and agent $j(0)$ announces (R, a). By Lemma C.1, all subgame perfect equilibrium outcomes of this subgame are top-ranked by agent $j(0)$. Thus, since agent $j(0)$ has a profitable deviation leading to a top-ranked outcome, this case is not an equilibrium.

2. There exists an $\overline{R} \in \mathcal{R}$ and $j \in N$ such that $R^i = \overline{R}$ for all $i \in N \setminus \{j\}$, and $b = f(\overline{R})$. There are two sub-cases to consider:

   (a) $j \neq j_d$. In this case, agent j_d has a deviation to trigger (and win) the integer game of case (i), by submitting $R^{j_d} = \perp$. Thus, either this case is not an equilibrium, or j_d already gets a top-ranked outcome.

   (b) $j = j_d$. If $j_d \neq j(0)$, then agent $j(0)$ has a profitable deviation yielding him a top-ranked outcome – namely, he triggers the integer game of case (i) by submitting $R^{j(0)} = \perp$. This deviation is profitable by (iii) of Condition α. Alternatively, if $j_d = j(0)$ then he can deviate to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. As in case 1 above, this situation is the same as the subgame of MR$'$ that follows the history h' in which all agents but $j(0)$ announce $(\overline{R}, b)$ and agent $j(0)$ announce (R, a). Again, by Lemma C.1, all subgame perfect equilibrium outcomes of this subgame are top-ranked by agent $j(0)$. Thus, since agent $j(0)$ has a profitable deviation leading to a top-ranked outcome, this case is not an equilibrium.

Thus, in all cases above, the equilibrium outcomes are always top-ranked by agent j_d. All other cases not included above are captured by case (iii) in Stage 0(b) of the protocol, in which j_d chooses an outcome. Thus, whenever the equilibrium profile $s(R)$ leads to an outcome $b \neq f(R)$, a deviation by an agent j_d would lead to a

subgame in which all SPE yield an outcome that is top-ranked by j_d. In particular, this means that there can be no such equilibrium s: Since $s(R)$ leads to an outcome $b \neq f(R)$, there must exists an agent j_d for whom b is not top-ranked (by NVP). Thus, j_d will deviate, leading to an outcome that is top-ranked by him. ■

Lemma C.12 *s^{MRP} is maximally-dispersed.*

Proof: Fix some $i \in N$, $R \in \mathcal{R}$, $h \in H \setminus Z$, and R^ψ-deviation $s_i' \neq s_i$ of agent i. Since the prescription of s^{MRP} is for every agent to play the same strategy at h and $N \geq 3$, when an agent deviates this deviation is noticeable. Furthermore, there is no $\overline{R}^\psi$ that leads to the same outcome under this deviation: That is, for every $\overline{R}^\psi \neq R^\psi$ it holds that $H\left(s_{-i} \circ s_i'\left(\overline{R}^\psi\right)|_h\right) \neq H\left(s_{-i} \circ s_i'\left(R^\psi\right)|_h\right)$. Thus,

$$\text{LD}_h\left(H\left(s_{-i} \circ s_i'\left(R^\psi\right)|_h\right), s\right) = \{R\}.$$

Now, in stage 0(a), it holds that

$$\text{L}\left(H\left(s\left(R^\psi\right)\right), s\right) = \text{LD}\left(H\left(s\left(R^\psi\right)\right), s\right) = \mathcal{R}|_a,$$

and so

$$\text{PT}\left(H\left(s\left(R^\psi\right)\right), s\right) = \mathcal{R}|_a.$$

Finally, in stages 0(b) and onwards agents always submit the action R as part of their strategies. No deviation from any $R' \neq R$ will lead to the same coordination on R. Thus,

$$\text{LD}_h\left(H\left(s\left(R^\psi\right)|_h\right), s\right) = \{R\}.$$

■

Lemma C.13 *If f satisfies $\alpha^{\max}$ and NVP, then* MRP *satisfies maximal deviability with respect to f.*

Proof: Fix some $R \in \mathcal{R}$ and a strategy profile s for which $g^{\text{MRP}}(H(s(R))) = b \neq f(R)$. Suppose first that in stage 0(a) of the mechanism it is *not* the case that $s_i(R, \epsilon) = b$ for all $i \in N$. Thus, at least $N - 1$ agents have a unilateral deviation that would trigger the integer game and yield them a top-ranked outcome. By NVP and the fact that $b \neq f(R)$, it must be the case that for at least one of these $N - 1$ agents b is not top-ranked. Thus, this agent has a profitable deviation that yields him a top-ranked outcome.

Suppose next that in stage 0(a) of the mechanism it holds that $s_i(R, \epsilon) = b$ for all $i \in N$, and consider stage 0(b) with a history $h = ((a_1, n_1), \ldots, (a_N, n_N))$. Suppose that for every $i \in N$ the strategy is $s(R, h) = (R^i, n^i)$. There are two cases to consider:

1. There exists an $\overline{R} \in \mathcal{R}$ such that $R^i = \overline{R}$ for all i, and $b = f(\overline{R})$. In this case, consider a deviation by agent $j(0)$ in the sequence $j(\overline{R}, R, b)$ to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. This situation is then the same as the subgame of MR$'$ that follows the history h' in which all agents but $j(0)$ announce $(\overline{R}, b)$ and agent $j(0)$ announces (R, a). By Lemma C.9, either $s(R)$ is a SPE of the subgame rooted at h', or some agent has a deviation that yields him a top-ranked outcome. In the latter case, we are done. For the former case, Lemma C.1 implies that all SPE outcomes of this subgame are top-ranked by agent $j(0)$. Thus, agent $j(0)$ has a profitable deviation at stage 0(b) that leads to a top-ranked outcome.

2. There exists an $\overline{R} \in \mathcal{R}$ and $j \in N$ such that $R^i = \overline{R}$ for all $i \in N \setminus \{j\}$, and $b = f(\overline{R})$. There are two sub-cases to consider:

   (a) $j \neq j_d$. In this case, agent j_d has a deviation to trigger (and win) the integer game of case (i), by submitting $R^{j_d} = \perp$. Thus, either agent j_d has a profitable deviation yielding him a top-ranked outcome, or j_d already gets a top-ranked outcome.

   (b) $j = j_d$. If $j_d \neq j(0)$, then agent $j(0)$ has a profitable deviation yielding him a top-ranked outcome – namely, he triggers the integer game of case (i) by submitting $R^{j(0)} = \perp$. This deviation is profitable by (iii) of Condition $\alpha^{\max}$. Alternatively, if $j_d = j(0)$ then he can deviate to $s_{j(0)}(h, R) = (R, 0)$. This leads to case (ii) of Stage 0(b), and a continuation to Stage 1. As in case 1 above, this situation is the same as the subgame of MR$'$ that follows the history h' in which all agents but $j(0)$ announce $(\overline{R}, b)$ and agent $j(0)$ announce (R, a). Again, either $s(R)$ is a SPE of the subgame rooted at h', in which case agent $j(0)$ has a profitable deviation at stage 0(b) that leads to a top-ranked outcome, or some agent has a deviation that yields him a top-ranked outcome.

Thus, in all cases above, either some agent has a deviation from $s(R)$ that leads to a top-ranked outcome, or the outcome following a deviation by agent j_d is top-ranked by agent j_d. All other cases not included above are captured by case (iii) in Stage 0(b) of the protocol, in which j_d chooses an outcome. Thus, whenever the profile $s(R)$ leads to an outcome $b \neq f(R)$, either some agent has a deviation yielding him a top-ranked outcome, or a deviation by an agent j_d would lead to a subgame in which

all outcomes are top-ranked by j_d. In particular, this means that, since $s(R)$ leads to an outcome $b \neq f(R)$, there must exists an agent j_d for whom b is not top-ranked (by NVP). Thus, j_d has a profitable deviation leading to an outcome that is top-ranked by him. ■

We are now ready to prove Theorems 5.2 and 5.5.

Proof of Theorem 5.2: There are two parts to the proof, corresponding to the two bullets of Definition 2.7. The first part of the proof follows. Since f can be implemented in subgame perfect equilibrium, by Abreu and Sen (1990) f must satisfy Condition α. By Lemma C.10, for every $R \in \mathcal{R}$ the strategy $s^{\text{MRP}}(R)$ is a SPE of MRP at R. Furthermore, by Lemma C.12, s^{MRP} is maximally-dispersed. Thus, by Lemma C.4, the strategy profile s^* satisfying $s_i^*(R^\psi, h) = s_i^{\text{MRP}}(R, h)$ for all i, R, and h is an information-sensitive subgame perfect equilibrium at ψ. Note that for all $R, \overline{R} \in \mathcal{R}$ such that $f(R) = f(\overline{R})$ it holds that $s^*(R^\psi, \epsilon) = s^*(\overline{R}^\psi, \epsilon)$, and that these both lead to terminal histories. In particular, this implies that $H(s^*(\mathcal{R}^\psi)) = H(s^*(\overline{R}^\psi))$, and so the strategy profile s^* satisfies bullets 1(a) and 1(b) of Definition 2.6 and bullet 1(c) of Definition 2.7.

Next, the second part of the proof is as follows. Let s be an information-sensitive subgame perfect equilibrium of MRP at ψ. Since ψ is lexicographic, Lemma C.5 implies that the profile s^o satisfying $s_i^o(R, h) = s_i(R^\psi, h)$ for all i, R, and h is a SPE of MRP. Now, since f satisfies NVP by assumption, Lemma C.11 implies that for every $R \in \mathcal{R}$, all SPE of MRP at R lead to the outcome $f(R)$. Since s and s^o lead to the same outcome, this implies that for every $R \in \mathcal{R}$, the profile $s(R^\psi)$ leads to the outcome $f(R)$. Thus, bullet 2 of Definition 2.6 is also satisfied.

Hence, the mechanism MRP is a privacy-protecting implementation of f at ψ. ■

Proof of Theorem 5.5: As in the proof of Theorem 5.2 there are two parts to the proof. The first part is identical to the proof of Theorem 5.2, implying the existence of a strategy profile s^* satisfying bullets 1(a) and 1(b) of Definition 2.6 and bullet 1(c) of Definition 2.7.

The second part of the proof is as follows. First note that, by Lemma C.13, the mechanism MRP satisfies maximal deviability with respect to f. Let s be an information-sensitive subgame perfect equilibrium of MRP at ψ. Since ψ satisfies MWR, Lemma C.7 implies that, for every $R \in \mathcal{R}$, the profile s^o satisfying $s_i^o(R, h) = s_i(R^\psi, h)$ for all i and h either leads to the outcome $f(R)$, or it is a SPE of MRP at

R. In the former case, since $s(R^\psi)$ and $s^o(R)$ lead to the same outcome, $s(R^\psi)$ leads to the outcome $f(R)$. For the latter case recall that f satisfies NVP by assumption, and so Lemma C.11 implies that for every $R \in \mathcal{R}$, all SPE of MRP at R lead to the outcome $f(R)$. Again, since s and s^o lead to the same outcome, this implies that for every $R \in \mathcal{R}$, the profile $s(R^\psi)$ leads to the outcome $f(R)$. Thus, bullet 2 of Definition 2.6 is also satisfied.

Hence, the mechanism MRP is a privacy-protecting implementation of f at ψ. ■

D Proofs from Section 6

D.1 Proof of Theorem 6.4

Proof: Suppose that there exists a normal-form mechanism (H, A, g) and a strategy profile s^* that satisfy bullets (b) and (c) of Definition 6.2. We will show that s^* is not an information-sensitive Nash equilibrium, and hence does not satisfy bullet (a), for a particular lexicographic ψ that we will construct.

Since $|\mathcal{R}| \geq 3$, there exists an outcome $a \in \{0, 1\}$ such that $|\mathcal{R}_a| \geq 2$, where $\mathcal{R}_a = \{R \in \mathcal{R} : f(R) = a\}$. This implies that there exists an agent i and a profile $R \in \mathcal{R}_a$ for which $(1-a)R_i a$. Now define ψ as follows: First, $(a, \{R\})P_i^\psi(a, \mathcal{R}_a)$. Next, if $aR_i(1-a)$, then $(1-a, \{R\})P_i^\psi(a, \mathcal{R}_a)$ and $(1-a, \{R_{1-a}\})P_i^\psi(a, \mathcal{R}_a)$. All other preferences are fixed in an arbitrary (lexicographic) way.

Suppose the true intrinsic preferences are R, and consider a deviation from s^* by agent i to s_i'. If either $g(H(s_{-i}^* \circ s_i'(R^\psi))) = a$ or $aR_i(1-a)$ and $H(s_{-i}^* \circ s_i'(R^\psi))$ is off the equilibrium path, then define $\text{PT}^\psi(H(s_{-i}^* \circ s_i'(R^\psi)), s^*) \stackrel{\text{def}}{=} \{R\}$, and note that this is one-deviation consistent. If $aR_i(1-a)$ and $H(s_{-i}^* \circ s_i'(R^\psi))$ is on the equilibrium path, then note that $g(H(s_{-i}^* \circ s_i'(R^\psi))) = 1-a$. In this case we must have $\text{PT}^\psi(H(s_{-i}^* \circ s_i'(R^\psi)), s^*) = \{\mathcal{R}_{1-a}\}$ by definition.

Now, we claim that with the constructed ψ and sets of possible types PT, the profile s^* is not an information-sensitive Nash equilibrium. Suppose the state is R^ψ and agent i deviates to s_i'. There are various possibilities: The outcome may be unchanged and remain a, in which case agent agent i benefits since the new and strictly preferred set of possible types is $\{R\}$. Alternatively, the outcome may be changed to $(1-a)$: If $(1-a)P_i a$ then since ψ is lexicographic agent i strictly benefits, and if $aR_i(1-a)$ then agent i benefits since the new and strictly preferred set of possible types is either $\{R\}$ or $\mathcal{R}_{1-a}$. Thus, in all cases agent i strictly benefits from

a deviation from s^* at R^ψ, and so s^* is not an equilibrium. ■

D.2 Proofs of Theorems 6.8 and 6.9

The mechanism used for Theorems 6.8 and 6.9 is the following variant of MRP from Section C.3. The mechanism is parametrized by a partition Π of $\mathcal{R}$. Label each element of Π by a unique number, and denote by $\pi(R)$ the label given to the element of Π that contains R.

Mechanism MRP^Π:

- **Stage 0(a):** Each agent i simultaneously submits a pair $(a_i, n_i, \pi_i) \in \mathcal{O} \times \mathbb{Z} \times \mathbb{N}$. If $a_1 = \ldots = a_N$, $\pi_1 = \ldots = \pi_N$, and $f(R) = a_1$ for all R satisfying $\pi(R) = \pi_1$, then the outcome is a_1. If there exists $j_d \in N$, $a \in \mathcal{O}$, and $\pi \in \mathbb{N}$ such that $a_i = a$ and $\pi_i = \pi$ for all agents $i \neq j_d$ but $a_{j_d} \neq a$, then go to Stage 0(b). Otherwise, the agent who submitted the highest integer n_i chooses the outcome of the mechanism.
- **Stage 0(b):** Suppose the history is $h = ((a_1, n_1, \pi_1), \ldots, (a_N, n_N, \pi_N))$. Continue as in MRP.
- **Stage k, $k = 1, \ldots, \ell$:** Continue as in MRP.

Consider the following strategy s^{MRP^Π}:

- In Stage 0(a), $s_i^{\text{MRP}^\Pi}(R, \epsilon) = (f(R), 0, \pi(R))$ for every $i \in N$.
- In Stage 0(b), profile R, and history h, every agent i submits $s_i^{\text{MRP}^\Pi}(R, h) = (R, 0)$.
- For all subsequent rounds (of the MR$'$ part of the mechanism), agents always announce $(0, R)$, as in $s^{\text{MR}'}$.

The proofs of Theorems 6.8 and 6.9 are nearly the same as those of Theorems 5.2 and 5.5. The only difference is in condition 1(c) of Definition 6.5. We need to show that if R and $\overline{R}$ are such that $\overline{R} \in \Pi(R)$, then $H(s^{\text{MRP}^\Pi}(\mathcal{R}^\psi)) = H(s^{\text{MRP}^\Pi}(\overline{R}^\psi))$. We also need to show that if R and $\overline{R}$ are such that $\overline{R} \notin \Pi(R)$, then $H(s^{\text{MRP}^\Pi}(\mathcal{R}^\psi)) \neq H(s^{\text{MRP}^\Pi}(\overline{R}^\psi))$. These are immediate from the specification of the strategy profile

s^{MRP^Π}: In the former case, the profile prescribes the same strategy, since in that case $\pi(R) = \pi(\overline{R})$. In the latter case the prescription is different, since in that case $\pi(R) \neq \pi(\overline{R})$ (and so the history differs in stage 0(a)).

References

Abreu, D. and Sen, A. (1990). Subgame perfect implementation: A necessary and almost sufficient condition. *Journal of Economic theory*, **50** 285–299.

Abreu, D. and Sen, A. (1991). Virtual implementation in nash equilibrium. *Econometrica: Journal of the Econometric Society* 997–1021.

Bernheim, B. (1994). A theory of conformity. *Journal of political Economy* 841–877.

Calzolari, G. and Pavan, A. (2006a). Monopoly with resale. *The RAND Journal of Economics*, **37** 362–375.

Calzolari, G. and Pavan, A. (2006b). On the optimality of privacy in sequential contracting. *Journal of Economic Theory*, **130** 168–204.

Chen, Y., Chong, S., Kash, I., Moran, T. and Vadhan, S. (2011). Truthful mechanisms for agents that value privacy. *arXiv preprint arXiv:1111.5472*.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 202–210.

Dutta, B. and Sen, A. (2012). Nash implementation with partially honest individuals. *Games and Economic Behavior*, **74** 154–169.

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, **1** 60–79.

Ghosh, A. and Roth, A. (2011). Selling privacy at auction. In *Proceedings of the 12th ACM conference on Electronic commerce*. ACM, 199–208.

Glazer, A. and Konrad, K. (1996). A signaling explanation for charity. *The American Economic Review*, **86** 1019–1028.

Glazer, J. and Rubinstein, A. (1998). Motives and implementation: On the design of mechanisms to elicit opinions. *Journal of Economic Theory*, **79** 157–173.

Ireland, N. (1994). On limiting the market for status signals. *Journal of public Economics*, **53** 91–110.

Jackson, M. (2001). A crash course in implementation theory. *Social choice and welfare*, **18** 655–708.

Maskin, E. (1999). Nash equilibrium and welfare optimality*. *The Review of Economic Studies*, **66** 23–38.

Maskin, E. and Sjöström, T. (2002). Implementation theory. *Handbook of social Choice and Welfare*, **1** 237–288.

Matsushima, H. (2008a). Behavioral aspects of implementation theory. *Economics Letters*, **100** 161–164.

Matsushima, H. (2008b). Role of honesty in full implementation. *Journal of Economic Theory*, **139** 353–359.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*. IEEE Computer Society, 94–103.

Miltersen, P., Nielsen, J. and Triandopoulos, N. (2009). Privacy-enhancing auctions using rational cryptography. *Advances in Cryptology-CRYPTO 2009* 541–558.

Moore, J. and Repullo, R. (1988). Subgame perfect implementation. *Econometrica: Journal of the Econometric Society* 1191–1220.

Muller, E. and Satterthwaite, M. (1977). The equivalence of strong positive association and strategy-proofness. *Journal of Economic Theory*, **14** 412–418.

Naor, M., Pinkas, B. and Sumner, R. (1999). Privacy preserving auctions and mechanism design. In *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, 129–139.

Nissim, K., Orlandi, C. and Smorodinsky, R. (2012). Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 774–789.

Palfrey, T. (2002). Implementation theory. *Handbook of Game Theory with Economic Applications*, **3** 2271–2326.

Pavan, A., Segal, I. and Toikka, J. (2012). Dynamic mechanism design.

Postlewaite, A. and Wettstein, D. (1989). Feasible and continuous implementation. *The Review of Economic Studies*, **56** 603–611.

Saijo, T. (1987). On constant maskin monotonic social choice functions. *Journal of Economic Theory*, **42** 382–386.

Serrano, R. (2004). The theory of implementation of social choice rules. *SIAM Review*, **46** 377–414.

Vartiainen, H. (2007). Subgame perfect implementation: A full characterization. *Journal of Economic Theory*, **133** 111–126.

Vohra, R. (2012). Dynamic mechanism design. *Surveys in Operations Research and Management Science*, **17** 60–68.

Xiao, D. (2011). Is privacy compatible with truthfulness. Tech. rep., Cryptology ePrint Archive, Report 2011/005.