MECHANISM DESIGN WITH PARTIALLY VERIFIABLE INFORMATION

By

Ronald Strausz

May 2016

COWLES FOUNDATION DISCUSSION PAPER NO. 2040

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Mechanism Design with Partially Verifiable Information

Roland Strausz*

May 23, 2016

Abstract

In mechanism design with (partially) verifiable information, the revelation principle obtains in full generality if allocations are modelled as the product set of outcomes and verifiable information. Incentive constraints fully characterize the implementable set of these product-allocations. The revelation principle does not generally hold when an allocation is modelled as only an outcome. However, any outcome of an implementable product-allocation is also implementable under this restricted modelling, provided that the mechanism designer can expand communication by adding unverifiable messages and restrict communication by limiting the use of messages. A canonical representation of such mechanisms is presented, implying that an inalienable right of the agent to withhold evidence does not affect implementability.

JEL Classification Numbers: D82
Keywords: Revelation principle, Mechanism Design, Verifiable Information

*Contact details: Humboldt-Universität zu Berlin, strauszr@wiwi.hu-berlin.de. This paper was written during my visist at the Cowles Foundation at Yale University in the spring of 2016. I thank Dirk Bergemann, Eduardo Faingold, Francoise Forges, Tibor Heumann, Johannes Hörner, Navin Kartik, Daniel Krähmer, Mallesh Pai, Juuso Toikka, Juuso Välimäki, and Alex Wolitzky for extremely helpful discussions and comments on earlier drafts.

1 Introduction

Focusing exclusively on the role of asymmetric information, mechanism design studies the extent to which the distribution of information restricts economic allocations. Ideally, the theory places no limitations on the ability of economic agents to interact and communicate, in principle allowing any type of game or mechanism to govern their communication and interactions.

The revelation principle plays a crucial role in enabling mechanism design to achieve, for a given information structure, its goal of analyzing unrestricted mechanisms. The principle is well understood in "standard mechanism design", which I define as a context in which economic agents can fully commit themselves to any possible mechanism but cannot credibly reveal their private information in any other form than by exchanging (unverifiable) messages.

For environments in which agents have (partially) verifiable information, the applicability of the revelation principle seems less well understood; even under full commitment. Following observations in Green and Laffont (1986), a general revelation principle for such settings is currently not available.[1] A failure of the revelation principle for these settings suggests that mechanism design with verifiability information differs fundamentally from mechanism design without verifiability.

The main goal of this paper is to argue that there is no such fundamental difference and that the confusion concerning the validity of the revelation principle is an artifact of the adopted modelling. More precisely, the tradition to model verifiable information as part of the communication rather than the economic allocation causes a failure of the revelation principle in general.

Hence, this paper's main departure from the existing literature on verifiability and mechanism design is that it models the presentation of the verified information as part of the economic allocation.[2] More specifically, it shows that by modeling the economic allocation as the product set of the pay-off relevant outcomes *and* the pay-off irrelevant provision of verifiable information (or evidence), the revelation principal obtains as usual. Moreover, any outcome associated with an implementable product-allocation of outcome and evidence is also implementable by mechanisms that limit the implementation to pay-off relevant outcomes only, under the provision that the following two elementary operations in the design of mechanisms are available: 1)

[1] E.g., Green and Laffont (1986), Bull and Watson (2007), and Deneckere and Severinov (2008) obtain some variants of the revelation principle only for certain subclasses of models.

[2] In the context of implementation without private information, Kartik and Tecieux (2012) make a similar point.

broadening communication by adding (non-verifiable) messages; and 2) restricting communication to a subset of available messages.

Because mechanism design (often implicitly) assumes the availability of both operations,[3] the failure of the revelation principle is not due to a fundamental difference of verifiability messages in mechanism design.[4] Regardless of whether information is partially verifiable, the revelation principle holds provided that verifiable information is modelled as being part of the implementable allocation. Moreover, even if one does not allow mechanisms to do so, the set of implementable outcomes which these allocations imply is still fully attainable if these restricted mechanisms allow the two rather elementary operations of broadening and restricting communication.

Instead, Green and Laffont (1986) model the presentation of verifiable information as hardwired restrictions on the agent's reporting structure rather than being part of the implementable allocation. The subsequent literature on mechanism design and verifiable information adopts this modeling approach as well. In principle, one can however just as well capture the verifiable information as part of the economic allocation.[5] It is however exactly this failure to do so that causes this literature's struggles with the validity of the revelation principle. A compelling reason for modeling the verifiable information as an explicit part of the economic allocation is therefore that it circumvents any conceptual problems concerning the revelation principle.[6]

Related literature

[3]E.g., when motivating the agent's restricted message sets on page 251, Green and Laffont themselves explicitly appeal to the principal's ability to restrict the agent's message space by imposing severe enough punishments. Deneckere and Severinov (2008) assumption of a "worst outcome" plays the same role; it acts as a severe enough punishment by which the principal restricts communication.

[4]In standard mechanism design, the first operation is clearly also essential for the revelation principle to hold, while the second operation is available naturally: the mechanism can always "prevent" any off-limit message by interpreting it as some specific available message and induce the same allocation. When there is no verifiable information, this does not affect the agent's reporting incentives so that preventing the agent to send additional messages is, effectively, implicit in the model. As Example 3 illustrates, this is not the case with verifiable information. Hence, a subtle but conceptual difference associated with the verifiability of information is the weaker ability to restrict communication when information is (partially) verifiable. See however the previous footnote concerning other natural features that, in mechanism design models with verifiable information, effectively imply the ability to restrict communication as well.

[5]E.g., Bull and Watson (2007, p. 79) explicitly mention this possibility but exclude it on the basis of pay-off relevance: "the players evidentiary actions are not directly payoff-relevant. Thus, we do not need to include evidence in the definition of the "outcome"."

[6]Section 5 shows that the subsequent set of implementable outcomes is also implementable when the agent's provision of evidence is modelled as an inalienable action.

Most points and observations presented in this paper have, in some way or another, also been raised in the existing literature of mechanism design with verifiable information or in the literature on (unique) implementation with perfect information. Indeed, most of this paper's findings can be understood as a reinterpretation of similar findings and notions there. In order to put the insights of this paper in perspective, it is therefore helpful to carefully point out its relations to previous results.

Green and Laffont (1986) were the first to note a failure of a general revelation principal in mechanism design problems with (partially) verifiable private information. They obtain a revelation principle only under a so-called *nested range condition*, where the agent's verifiability exhibits a nested structure. They show by explicit examples that without this condition, the revelation principle in general fails. They note that this failure limits the applicability of mechanism design to study settings with partially verifiable information, because one cannot characterize the set of implementable allocations. Green and Laffont do not model the presentation of evidence as part of the economic allocation and, the two examples below clarify that they implicitly restrict the design of mechanisms.

Singh and Wittman (2001) give plausible examples of concrete economic environments for which the nested range condition of Green and Laffont is violated. For principal-agent models that satisfy a *unanimity* condition on the agent's preferences, they derive necessary and sufficient conditions for the implementability of a social choice function regardless of the underlying verifiability structure. The authors do not discuss possible extensions of direct mechanisms such as broadening and restricting communication. They also do not model the presentation of evidence as part of the economic allocation.

Bull and Watson (2007) explicitly address the validity of the revelation principle in mechanism design with partially verifiable information and multiple agents. They do not model the provision of evidence as part of the economic allocation and motivate this exclusion by the fact that the provision of evidence is not directly payoff-relevant. In addition, their mechanism design setup does not allow the operation of restricting communication. Due to both restrictions, a general revelation principle in their context is not available, but the authors show that the principle obtains under an *evidentiary normality* condition, which is closely related to the nested range condition of Green and Laffont (1986).

Also Deneckere and Severinov (2008) report a failure of the usual revelation principle in mechanism design with partially verifiable information. They, instead, present an extended revelation principle, which uses dynamic mechanisms. They further refine the concepts of nested information under which the revelation principle holds and

point to similar notions in Postlewaite and Schmeidler (1986) and Lipman and Seppi (1995). While Deneckere and Severinov do not model the presentation of evidence as part of the economic allocation, they, however, explicitly allow the operation of adding non-verifiable messages, which they define as "cheap talk". Moreover, because the authors mostly focus on principal-agent problems for which there is a type-independent "worst outcome" for the agent, they also implicitly allow the principal to restrict the agent's communication, because by committing to implement the worst outcome for a certain message, the principal can ensure that the agent will never use this message. Deneckere and Severinov moreover express a direct interest in analyzing the effect of ad hoc limitations on the principal's ability to design communication rules such as limitations on the number of messages which agent can send.

The literature on (unique) implementation with perfect information has also studied verifiable evidence (e.g. Bull and Watson, 2004, Ben-Porath and Lipman, 2012 and Kartik and Tercieux, 2012). From the perspective of this literature, the idea of extending the outcome space as presented in this paper, is not new. In particular, Section 4 in Kartik and Tercieux (2012) consider the same kind of extended allocation space and also show that restricting to mechanisms that consider the agent's evidence provision as an inalienable action does not reduce the set of implementable outcomes.[7]

Analyzing the role of verifiable information in a game theoretical rather than a mechanism design context, Forges and Koessler (2005) study communication between players with private but partially verifiable information. Since the authors do not follow a mechanism design perspective, they do not use the notion of mechanisms as implementing economic allocations. Yet, the revelation principles they obtain and their underlying proofs are closely linked to the one shown in this paper. Importantly, the authors also explicitly point out the importance of broadening and restricting communication for expanding the set of equilibrium outcomes in their game theoretical framework.

In addition to Forges and Koessler (2005), the revelation of verifiable information in games with incomplete information is studied in, for instance, Hagenbach et al. (2014) and the extensive literature on (Bayesian) persuasion (e.g. Glazer and Rubinstein, 2004 and Kamenica and Gentzkow, 2011). The main difference to this literature is that players cannot commit (all) their actions to a mechanism.

[7]In a private communication, the authors sent notes in which they derive the counterpart of my Propositions 3 and 4 in a mechanism design context with quasi-linearity and transfers.

2 The Green and Laffont (1986) example

This section first reiterates the example by which Green and Laffont (1986) demonstrate the failure of the revelation principle and, subsequently, illustrates how to recover it by a natural reinterpretation of an economic allocation.

Example 1: Green and Laffont (1986)
Consider a principal and one agent, who can be of three types $\Theta_1 = \{\theta_1, \theta_2, \theta_3\}$. The set of outcomes is $X_1 = \{x_1, x_2\}$. The agent has partially verifiable information, which Green and Laffont concisely capture by type-specific message sets $M(\theta_i)$ with the interpretation that type θ_i can only send messages from the set $M(\theta_i)$. In their specific example they consider the sets $M_1(\theta_1) = \{\theta_1, \theta_2\}$, $M_1(\theta_2) = \{\theta_2, \theta_3\}$, $M_1(\theta_3) = \{\theta_3\}$. The agent's utilities $u_1(x, \theta)$ are as follows:

$u_1(x,\theta)$	θ_1	θ_2	θ_3
x_1	10	5	10
x_2	15	10	15

For this example, Green and Laffont show that the direct mechanism $g_1 : \Theta \to X$ with $g_1(\theta_1) = g_1(\theta_2) = x_1$ and $g_1(\theta_3) = x_2$ induces a game that implements the social choice function $f_1(\theta_1) = x_1$, $f_1(\theta_2) = f_1(\theta_3) = x_2$. This is so, because type θ_1, who cannot send the message θ_3, optimally sends the message θ_1, which results in $x_1 = f_1(\theta_1)$. Type θ_2, who cannot send the message θ_1, optimally sends the message θ_3, which results in $x_2 = f_1(\theta_2)$. Type θ_3, who can only send the message θ_3, optimally sends the message θ_3, which results in $x_2 = f_1(\theta_3)$.

Green and Laffont observe that while direct, the mechanism is not truthful, because it induces type θ_2 to misreport his type as θ_3. They, subsequently, establish a failure of the revelation principle, because a truthful direct mechanism $\hat{g}_1$ that implements f_1, requires $\hat{g}_1(\theta_1) = x_1$, $\hat{g}_1(\theta_2) = x_2$, $\hat{g}_1(\theta_3) = x_2$. This mechanism is however not incentive compatible, because it induces type θ_1 to report θ_2.

We can however implement the social choice function f_1 with a truthful direct mechanism if we extend the concept of an allocation as follows. In addition to an outcome $x \in X$, an allocation also describes a verifiable message $\theta \in \Theta$ which the agent is to send. Hence, let the set $Y = X \times \Theta$ represents this extended set of allocations with a typical element $y = (x, \theta) \in Y$. Define utilities as follows:

$$\hat{u}(x, \theta) = \begin{cases} u(x, \theta') & \text{if } \theta \in M(\theta') \\ -\infty & \text{otherwise.} \end{cases}$$

In this extended context, a direct mechanism is a function $\tilde{y} = (\tilde{x}, \tilde{\theta}) : \Theta \to Y$ from the set of non-verifiable claims about Θ to the extended set of allocations of outcomes X and verifiable messages about Θ.[8] Its interpretation is that if the agent send the non-verifiable claim θ_i, the mechanism picks $\tilde{x}(\theta_i) \in X$ and the agent must present the message $\tilde{\theta}(\theta_i) \in \Theta$. The direct mechanism $y(\theta_1) = (x_1, \theta_1)$, $y(\theta_2) = (x_2, \theta_3)$, $y(\theta_3) = (x_2, \theta_3)$ is incentive compatible (truthful) and implements the allocations in X as intended by the social choice function f_1.

3 The Mechanism Design Setup

The above example suggests that by extending the concept of an implementable allocation, one can recover the revelation principle. This section make precise the sense in which this insight is general. It is most instructive to do so in the original framework of Green and Laffont (1986), because it is the simplest framework to illustrate the two additional requirements for this extension to work: The ability to both extend and limit communication.

Hence, consider a principal facing an agent with utility function $u(x, \theta)$, which depends on a characteristic $\theta \in \Theta$ and an outcome $x \in X$. For concreteness, we assume that both sets are finite: $\Theta = \{\theta_1, \ldots, \theta_K\}$ and $X = \{x_1, \ldots, x_L\}$ with $K, L \in \mathbb{N}$.[9] The agent knows θ, whereas the principal only knows that $\theta \in \Theta$. The agent has verifiable information represented by a correspondence $M : \Theta \to \Theta$ with the interpretation that type θ_i can only send messages about θ from the set $M(\theta_i)$. Hence, a type θ describes both the agent's preferences over X and an available message set. In short, we can represent the principal-agent problem of Green and Laffont by a structure $\Gamma = \{\Theta, X, M(\cdot), u(\cdot, \cdot)\}$, which consists and describes all the primitives of the principal-agent model.

Fully in line with the usual goal of mechanism design, Green and Laffont (p.448) state their intention to "study the class of social choice functions f from Θ into X

[8] Hence, claims and messages are different objects in this context and not synonyms. While we will use $\hat{u}$ and the mechanism $\tilde{y}$ only as hypothetical constructs for deriving a revelation principle, they allow the following literal interpretation. Although an agent can costlessly make any unverifiable claim about his type, he has a prohibitively high cost to back up his claim if he cannot present the verifiable information to substantiate it. Hence, a person with only \$10 dollars in his pocket, can claim he has \$20, but has a prohibitively high cost of actually retrieving \$20 from his pocket. In contrast, a person with \$20 dollars in his pocket, can claim to have \$20 dollars and also produce the \$20 at zero costs.

[9] All arguments naturally extend if Θ and X are subsets of some more general Euclidean spaces.

that can be achieved despite the asymmetry of information between the two players." For this, they define a direct mechanism as follows.

Definition 1: A mechanism $(M(\cdot), g)$ consists of a correspondence $M : \Theta \to \Theta$ such that $\theta \in M(\theta)$ for all $\theta \in \Theta$, and an outcome function $g : \Theta \to X$.

Hence, the mechanism $(M(\cdot), g)$ presents the agent with a single-person decision problem in which an agent of type θ has to pick some θ but in which his choice is restricted to his message set $M(\theta)$. Following Green and Laffont, we can describe the agent's optimal decision behavior as follows. Given the correspondence $M(\cdot)$, the outcome function g induces a *response rule* $\phi_g : \Theta \to \Theta$ defined by[10]

$$\phi_g(\theta) \in \arg \max_{m \in M(\theta)} u(g(m), \theta).$$

This leads to the following two notions of implementability.

Definition 2: A social choice function $f : \Theta \to X$ is $M(\cdot)$-*implementable* iff there exists an outcome function $g : \Theta \to X$ such that:

$$g(\phi_g(\theta)) = f(\theta) \text{ for any } \theta \text{ in } \Theta,$$

where $\phi_g(\cdot)$ is an induced response rule.

Definition 3: A social choice function $f : \Theta \to X$ is *truthfully* $M(\cdot)$-*implementable* iff there exists an outcome function $g^* : \Theta \to X$ such that:

$$g^*(\phi_{g^*}(\theta)) = f(\theta) \text{ for any } \theta \text{ in } \Theta$$

and

$$\phi_{g^*}(\theta) = \theta.$$

The example in the previous section proves that there exists social choice functions that are $M(\cdot)$-implementable but not truthfully $M(\cdot)$-implementable. In this result, Green and Laffont see a failure of the revelation principle and the ensuing problem that one cannot, in general, characterize the set of implementable social choice functions for all principal-agent problems Γ.[11]

[10] Because Θ is finite, the maximum exists.

[11] Note that the agent's decision problem involves a type-dependent action set. Hence, extending this approach to multiple agents leads to the concern that the game induced by the mechanism does not, strictly speaking, correspond to a Bayesian Game. In the definitions following Harsanyi (1967), games with imperfect information require that the agent's action sets are type-independent. (See footnote 13 for more details and also Bull and Watson (p. 80, 2007) who point out that their "disclosure game [...] is a Bayesian game with type-contingent restrictions on actions".)

The next two examples suggest, however, that not only the notion of *truthfully* $M(\cdot)$-implementability is problematic, but that the more primitive notion of $M(\cdot)$-implementability also raises questions. In Example 2, the specified social choice function is not $M(\cdot)$-implementable, whereas it is implementable if the mechanism can, in addition to the messages in $M(.)$, also condition on two non-verifiable messages. In Example 3, the specified social choice function is not $M(\cdot)$-implementable, whereas it is implementable if the mechanism can limit the messages that can be sent.

Example 2: Too few messages
Consider a third outcome x_3 by duplicating outcome x_2 in the sense that each type θ is indifferent between x_3 and x_2. Hence, the set of outcomes is $X_2 = \{x_1, x_2, x_3\}$ with the utility

$u_2(x,\theta)$	θ_1	θ_2	θ_3
x_1	10	5	10
x_2	15	10	15
x_3	15	10	15

Suppose we want to implement the social choice function $f_2(\theta_1) = x_1$, $f_2(\theta_2) = x_2$, $f_2(\theta_3) = x_3$. Then, based on the reasoning in Example 1, it is straightforward to see that this social choice function is not $M(\cdot)$-implementable, but it is implementable by a mechanism that, in addition to reporting θ, asks for some extra cheap talk message $\hat{m} \in \hat{M} = \{a, b\}$ as follows

$$g_2(\theta, \hat{m}) = \begin{cases} x_2 & \text{if } (\theta, \hat{m}) = (\theta_3, a) \\ x_3 & \text{if } (\theta, \hat{m}) = (\theta_3, b) \\ x_1 & \text{otherwise.} \end{cases}$$

With the concept of an extended allocation as introduced in Example 1, the incentive compatible direct mechanism $y_2(\theta_1) = (x_1, \theta_1)$, $y_2(\theta_2) = (x_2, \theta_3)$, $y_2(\theta_3) = (x_3, \theta_3)$ implements the outcomes in X_2 as intended by the social choice function f_2. □

Example 3: Too many messages
Consider three type $\Theta_3 = \{\theta_1, \theta_2, \theta_3\}$ with two outcomes $X_3 = \{x_1, x_2\}$, message sets $M_3(\theta_1) = \{\theta_1, \theta_2\}$, $M_3(\theta_2) = \{\theta_2, \theta_3\}$, $M_3(\theta_3) = \{\theta_3\}$, and utilities

$u_3(x,\theta)$	θ_1	θ_2	θ_3
x_1	0	1	1
x_2	1	0	0

Consider the social choice function $f_3(\theta_1) = x_1$, $f_3(\theta_2) = f_3(\theta_3) = x_2$, inducing a utility 0 for each type. This social choice function is not $M(.)$-implementable. For suppose it is $M(.)$-implementable by some function $g_3 : \Theta \to X$. There are two cases for $g_3(\theta_2)$. Case 1: $g_3(\theta_2) = x_1$, but then type θ_2 can guarantee himself 1 by sending the message θ_2, which contradicts that he is supposed to get 0 under f_3. Case 2: $g_3(\theta_2) = x_2$, but then type θ_1 can guarantee himself 1 by sending the message θ_2, contradicting that he is supposed to get 0 under f_3. Note however that by restricting the mechanism to only messages $\{\theta_1, \theta_3\} \subset \Theta$ and setting $g_3(\theta_1) = x_1$ and $g_3(\theta_3) = x_2$, the social choice function f_3 is implementable. Hence, to implement f_3 it is crucial that the agent's communication is restricted: he is not allowed to send the message θ_2. Because Green and Laffont define a mechanism as consisting of an outcome function whose domain is the entire set of types Θ, they formally do not allow such restrictions in their framework. □

4 A Revelation Principle

The two last examples of the previous section suggest that for studying the restrictions on implementable outcomes, the notion of a direct mechanism is too restrictive. From a perspective of standard mechanism design, this seems a puzzling observation. This section argues however that the observation is more due to a restrictive modelling of an economic allocation rather than some deep fundamental difference to mechanism design with verifiable information. In particular, the revelation principle obtains as usual if an implementable allocation consists not only of the outcome x, but also the verifiable message m that is sent.

In order to see that such a definition of an allocation is, from a mechanism design perspective, also the more natural one, recall that mechanism design is primarily interested in the equilibrium outcomes which a mechanism induces rather than the specifics of the mechanism itself. More precisely, it is not the complete equilibrium outcome that matters, but only the "components" of these equilibrium outcomes that are relevant to the mechanism design problem. Hence, in order to model some environment with private information as a problem of mechanism design, it is crucial to identify first its relevant components.

Clearly, the direct pay-off relevant part, the outcome $x \in X$, is such a component. Yet, it is important to recognize that, because the verifiable messages allow the principal to screen among different types of agents, they are just as relevant. Put differently, verifiable messages are screening variables and because screening variables

are a crucial part of the mechanism design problem, they should be included as part of the implementable allocation.

In contrast, the tradition in the literature on mechanism design with verifiable information includes only the outcome $x \in X$ as part of the implementable allocation. Its motivation for not including the verifiable messages themselves lies in the suggestion that sending a verifiable message is costless and, thus, pay-off irrelevant. This view is however partially misguided, because whilst it is true that for a type who can send a verifiable message, the cost is zero, the cost is effectively infinite for a type who cannot send it. Hence, one can just as well argue that, in contrast to unverifiable messages, verifiable messages actually exhibit an extreme form of payoff-relevance. Taking a pure pay-off relevance perspective, they should therefore be part of the implementable allocation.

Importantly, the inclusion of verifiable messages as part of the implementable allocation fully restores the revelation principle in mechanism design with verifiable information. In order to show this formally in the original context of Green and Laffont, one first has to make a modeling choice about an aspect, which Green and Laffont leave unspecified: whether the verifiable messages which the agent can send are "exhaustive".

Subsequent literature provides two perspectives on this. By interpreting that the agent's verifiable message is effectively a collection of possible pieces of evidence, Bull and Watson (2007) present a micro-foundation for the underlying verifiable messages, which implies that the agent can only send one verifiable message. In contrast, Deneckere and Severinov (2008) do not model this intermediate step of differentiating between the verifiable message and its underlying pieces of evidence. As the authors explicitly explain, without this distinction, it is appropriate to model the possibility that the agent can send multiple verifiable messages.

For the validity of the revelation principle it is inconsequential which modeling choice to make. For the exposition of the result, it is however less cumbersome to follow the interpretation of Bull and Watson (2007). In this case, the verifiable-message component of an equilibrium outcome that is induced by some mechanism is an element of Θ, whereas under the interpretation of Deneckere and Severinov (2008), where the agent could send multiple verifiable messages, a verifiable message is an element of the power set 2^{Θ}.

Given a principal-agent problem $\Gamma = \{\Theta, X, M, u\}$, define the *extended allocation set* $\hat{X} \equiv X \times \Theta$ and the *extended utility function* $\hat{u}(\hat{x}|\tilde{\theta}) = \hat{u}(x, \theta|\tilde{\theta})$ as

$$\hat{u}(x, \theta|\tilde{\theta}) \equiv \begin{cases} u(x, \tilde{\theta}) & \text{if } \theta \in M(\tilde{\theta}) \\ u(x, \tilde{\theta}) - C & \text{otherwise.} \end{cases} \tag{1}$$

with $C = \max_{x,\theta,x',\theta'} u(x,\theta) - u(x',\theta')$.[12,13] We can interpret the expanded structure $\hat{\Gamma} = \{\Theta, \hat{X}, M, \hat{u}\}$ as representing a principal-agent problem in which the principal wants to implement an *extended social choice function* $\hat{f} : \Theta \to \hat{X}$ given that the agent is privately informed about his type θ.

A social choice function $\hat{f}$ is implementable if there exists some single-person decision problem in which for any type θ there exists an optimal decision inducing the allocation $\hat{f}(\theta)$. A special class of such decision problems are incentive compatible direct mechanisms defined as follows.

Definition $\hat{1}$: An *incentive compatible direct mechanism* in $\hat{\Gamma}$ is a composite function $\hat{g} = (\hat{g}^1, \hat{g}^2)$ with $\hat{g}^1 : \Theta \to X$ and $\hat{g}^2 : \Theta \to \Theta$ such that

$$\hat{u}(\hat{g}(\theta)|\theta) \geq \hat{u}(\hat{g}(\theta')|\theta) \text{ for any } \theta, \theta' \in \Theta. \tag{2}$$

Hence, an incentive compatible direct mechanism $\hat{g}$ represents a single-person decision problem in which it is an optimal decision for the agent to report his type truthfully. We adapt Definition 2 to $\hat{\Gamma}$ as follows.

Definition $\hat{2}$: A social choice function $\hat{f} : \Theta \to \hat{X}$ is *$\hat{g}$-implementable* iff the direct mechanism $\hat{g} = \hat{f}$ is incentive compatible.

Standard arguments yield the revelation principle for the principal-agent problem $\hat{\Gamma}$: If there exists some single-person decision problem in which for any type θ there exists an optimal decision leading to the extended allocation $\hat{f}(\theta)$, then there exists an incentive compatible direct mechanism with $\hat{g}(\theta) = \hat{f}(\theta)$. Hence, the mechanism $\hat{g}$ implements the social choice function $\hat{f}$. Therefore the next proposition follows.

Proposition 1 (Revelation principle) *Any extended allocation $\hat{f}(\theta)$ that is the outcome of some single-agent decision problem in $\hat{\Gamma}$ is $\hat{g}$-implementable.*

While the previous proposition establishes a revelation principle for the principal-agent problem $\hat{\Gamma}$, it leaves open its relation to the underlying problem Γ.

[12]Because Θ and X are finite, C is well-defined. If the sets Θ and X are infinite, one may take the supremum rather than the maximum, which is well-defined provided that u is bounded. If u is unbounded, all arguments still go through by picking, for a given social welfare function f, a large enough (finite) value for C.

[13]Extending the agent's payoff function by introducing the prohibitively high cost C effectively renders the agent's action set type-independent and solves the issue pointed out in footnote 11 . It is an illustration of the idea of Harsanyi (p. 168, 1967) that "the assumption that a given strategy $s_i = s_i^0$ is not *available* to player i is equivalent, from a game-theoretical point of view, to the assumption that player i will never actually *use* strategy s [emphasis in the original]."

Proposition 2 *Consider some principal-agent problem* Γ *and its corresponding extension* $\hat{\Gamma}$*. If there exists some mechanism in* Γ *which implements the social choice function* $f : \Theta \to X$*, then there exists a function* $\hat{\theta} : \Theta \to \Theta$ *such that the extended social choice function* $\hat{f}(\cdot) = (f(\cdot), \hat{\theta}(\cdot))$ *is* $\hat{g}$*-implementable.*

Proof of Proposition 2: Suppose some decision problem implements the social choice function f in Γ. Then for type θ, some decision(s) leading to the outcome $f(\theta)$ and some verifiable message $\hat{\theta}(\theta) \in M(\theta)$ that he sends when achieving outcome $f(\theta)$ is optimal. Consider the direct mechanism $\hat{g} : \Theta \to \hat{X}$ with $\hat{g}^1(\theta) = f(\theta) \in X$ and $\hat{g}^2(\theta) = \hat{\theta}(\theta) \in M(\theta) \subset \Theta$. Fix some $\theta \in \Theta$. Inequality (2) holds for any θ' s.t. $\hat{\theta}(\theta') \notin M(\theta)$, because $u(f(\theta'), \theta) - C \leq \min_{x,\tilde{\theta}} u(x, \tilde{\theta}) \leq \hat{u}(f(\theta), \hat{\theta}(\theta))$, since $\hat{\theta}(\theta) \in M(\theta)$. Moreover, the optimality of the decision(s) leading to $f(\theta)$ and message $\hat{\theta}(\theta)$ imply that inequality (2) holds for any θ' s.t. $\hat{\theta}(\theta') \in M(\theta)$. It then follows that the constructed $\hat{g}$ satisfies (2) for any $\theta, \theta' \in \Theta$ so that $\hat{g}$ is an incentive compatible direct mechanism in $\hat{\Gamma}$. Hence $\hat{f}$ is $\hat{g}$-implementable. Q.E.D.

The main insight is therefore that, despite the presence of (partially) verifiable information, there is nothing peculiar about the principal-agent problem if we, appropriately, specify the concept of an implementable allocation. We can then use the revelation principle as usual to analyze the class of implementable allocations for all possible mechanisms. In particular, the incentive constraints (2) full characterize the set of implementable social choice functions $\hat{f}$. Taking the first component of $\hat{f}$ gives us the set of implementable outcomes $x \in X$.

5 Inalienable Mechanisms

Proposition 2 provides the answer to the question how to characterize the set implementable social choice functions f for any Γ: First characterize the set of implementable social choice functions $\hat{f}$ in the corresponding problem $\hat{\Gamma}$ by the incentive constraints (2). The set all implementable social choice function f can then be obtained in a second step by taking the first component of each implementable $\hat{f}$.

The procedure characterizes the set of implementable outcome via mechanisms $\hat{g} = (\hat{g}^1, \hat{g}^2)$ that map into the extended allocation $X \times \Theta$ rather than the outcome space X. One interpretation of such mechanisms is that they are *evidence-conditional*. The agent receives the allocation $\hat{g}^1(\theta) \in X$ conditional on presenting the evidence $\hat{g}^2(\theta) \in \Theta$. While the previous section shows that evidence-conditional mechanisms allow us to characterize the set of implementable outcomes with the usual tools of

mechanism design and, in particular, the notion of incentive compatible direct mechanisms, one may object that these mechanisms may be too coercive for some practical environments, because they effectively force the agent to present his evidence.[14]

This has lead the mechanism design literature with verifiability to introduce the concept of *inalienability*, which is the notion that the principal should not be able to force the agent to present any evidence, if the agent does not want to. The evidence-conditional mechanisms $\hat{g}$ have the strong flavor that they violate this notion, because the agent is forced to give up the evidence $\hat{g}^2(\theta)$ if he wants the allocation $\hat{g}^1(\theta)$. Following the concern that evidence-conditional mechanisms violate the notion of inalienability, we introduce the following definition:

Definition: An inalienable mechanism (M, g) consists of a set M and an outcome function $g : M \to X$.

Hence, the restrictive notion of an inalienable mechanism is that, just as in the original framework of Green and Laffont, it maps into the set of outcomes X instead of some larger set. While clearly relevant from a practical perspective, it is however important to be aware that, from a pure mechanism design perspective, the introduction of inalienability is a restriction on the choice of mechanisms that is not directly related to the presence of asymmetric information. Clearly, any additional restrictions on the available mechanisms can only reduce the set of implementable allocations.

Taking the concern of inalienability seriously, an interesting question is however to ask how restrictive the introduction of such mechanisms is in terms of implementability. Ideally, the set of implementable allocation with inalienable mechanisms coincides with the set of implementable allocations via unrestricted mechanisms.

One example of an inalienable mechanism is a direct mechanism, $g : \Theta \to X$, as modelled in Green and Laffont. Yet, Examples 2 and 3 show that the set of implementable allocations via these mechanisms is strictly smaller than the set of implementable allocations via unrestricted mechanisms. Indeed, starting with a direct mechanism of Green and Laffont and broadening it by adding unverifiable messages or reducing it further by restricting communication, still yields a mechanism that is inalienable. Hence, the examples show that, in general, there exist inalienable mechanisms which can implement outcomes that are not implementable via a direct mechanism in the sense of Green and Laffont.

The remainder of this section shows that starting with a direct mechanism of Green and Laffont and using the two elementary operations of broadening it by

[14] E.g., the "right to remain silent" is a right recognized in many of the world's legal systems.

adding unverifiable messages and limiting it by restricting communication, obtains an inalienable mechanism that implements an outcome that is implementable by some non-inalienable evidence-conditional mechanism $\hat{g}$. This result implies that inalienable mechanisms are not restrictive in terms of the implementable outcomes they induce. Because the proof is constructive, it shows exactly how to obtain the inalienable mechanism that implements the same outcome as its evidence-conditional counterpart $\hat{g}$.

Given a principal-agent problem Γ with the message correspondence $M(.)$, first extend the agent's message as follows:

$$\hat{M}(\theta) = M(\theta) \times \Theta.$$

As before an interpretation of this extension is that, in addition to a message from $M(\theta)$, an agent of type θ also makes some non-verifiable claim about his type Θ. However, this additional non-verifiable message does not necessarily be a literal claim about the agent's type. Another interpretation is that the agent has to say some natural number between 1 and K, which, given that there are K types, effectively is like reporting some Θ.

Let $\hat{M} \equiv \cup_{\theta \in \Theta} \hat{M}(\theta)$ and define a mechanism as follows:

Definition $\bar{1}$: A mechanism $(\bar{M}, \bar{g})$ in Γ consists of a set $\bar{M} \subseteq \hat{M}$ such that $\bar{M} \cap \hat{M}(\theta) \neq \emptyset$ for all $\theta \in \Theta$ and an outcome function $\bar{g} : \bar{M} \to X$.

Hence, a mechanism $(\bar{M}, \bar{g})$ is inalienable. Moreover, it is constructed by starting with the message sets $M(\theta)$, which Green and Laffont consider as primitives of the underlying principal-agent problem, extending them by adding non-verifiable messages, in the form of the set Θ, to obtain the extended messages set $\hat{M}$, and, subsequently, restricting this overall message set to $\bar{M}$, which is a (possibly strict) subset of $\hat{M}$. Hence, if, in the mechanism design problem, the principal has the ability to perform the two elementary operations of adding non-verifiable messages and restricting the agent's communication, then it is compelling that the principal can use mechanisms as defined in Definition $\bar{1}$.

As before, a mechanism $(\bar{M}, \bar{g})$ presents the agent of type θ with a single-person decision problem in which he has to pick some m from the message set $\bar{M}$ that is consistent with his message set $\hat{M}(\theta)$. That is, the mechanism induces a *response rule* $\bar{\phi}_{\bar{g}} : \Theta \to \bar{M}$ defined by[15]

$$\bar{\phi}_{\bar{g}}(\theta) \in \arg \max_{m \in \hat{M}(\theta) \cap \bar{M}} u(\bar{g}(m), \theta).$$

[15]The provision in Definition $\bar{1}$ that $\hat{M}(\theta) \cap \bar{M} \neq \emptyset$ for all $\theta \in \Theta$ implies that the agent does not maximize over an empty set.

Because the function $\bar{\phi}_{\bar{g}}$ maps Θ into the Cartesian product $\Theta \times \Theta$, it is convenient to write the composed function $\bar{\phi}_{\bar{g}}$ component-wise as $\bar{\phi}_{\bar{g}} = (\bar{\phi}^1_{\bar{g}}, \bar{\phi}^2_{\bar{g}})$ of two functions $\bar{\phi}^1_{\bar{g}} : \Theta \to \Theta$ and $\bar{\phi}^2_{\bar{g}} : \Theta \to \Theta$.

The adapted notions of a mechanism and a response rule lead to the following concept of implementability.

Definition $\bar{2}$: A social choice function $f : \Theta \to X$ is *$\bar{M}$-implementable* in Γ iff there exists a mechanism $(\bar{M}, \bar{g})$ with an outcome function $\bar{g}$ such that:

$$\bar{g}(\bar{\phi}_{\bar{g}}(\theta)) = f(\theta) \text{ for any } \theta \text{ in } \Theta, \tag{3}$$

where $\bar{\phi}_{\bar{g}}(\cdot)$ is a response rule with respect to the mechanism $(\bar{M}, \bar{g})$.

The next proposition makes precise the idea that any implementable outcome is implementable by an inalienable mechanism $(\bar{M}, \bar{g})$.

Proposition 3 *Consider a principal-agent problem Γ and the corresponding problem $\hat{\Gamma}$. If $\hat{f} = (\hat{x}, \hat{\theta})$ is $\hat{g}$-implementable in $\hat{\Gamma}$ and $\hat{\theta}(\theta) \in M(\theta)$ for all $\theta \in \Theta$, then $f = \hat{x}$ is $\bar{M}$-implementable in Γ.*

Proof of Proposition 3: Fix $\hat{f} = (\hat{x}, \hat{\theta})$ and define $\bar{M}$ as

$$\bar{M} = \{(\hat{\theta}(\theta_i), \theta_i) : \theta_i \in \Theta\}.$$

Because $\hat{\theta}(\theta) \in M(\theta)$, it holds by construction of $\hat{M}(\theta)$ that $\bar{M} \subset \cup_\theta \hat{M}(\theta)$. Define the outcome function $\bar{g} : \bar{M} \to X$ as $\bar{g}(\hat{\theta}(\theta), \theta) = \hat{x}(\theta)$ for any $(\hat{\theta}(\theta), \theta) \in \bar{M}$. Note $\hat{M}(\theta_i) \cap \bar{M} = (\hat{\theta}(\theta_i), i) \neq \emptyset$ for any $\theta_i \in \Theta$. Hence, $(\bar{M}, \bar{g})$ is a mechanism according to Definition $\bar{1}$. Moreover, because $\hat{M}(\theta_i) \cap \bar{M} = (\hat{\theta}(\theta_i), \theta_i)$ is a singleton, $(\hat{\theta}(\theta), \theta)$ is a response rule with respect to the mechanism $(\bar{M}, \bar{g})$. Hence, $\bar{\phi}_{\bar{g}}(\theta) = (\hat{\theta}(\theta), \theta)$ so that $\bar{g}(\bar{\phi}_{\bar{g}}(\theta)) = \bar{g}(\hat{\theta}(\theta), \theta) = \hat{x}(\theta) = f(\theta)$. Therefore, $f = \hat{x}$ is $\bar{M}$-implementable. Q.E.D.

By constructively deriving the incentive compatible mechanism $\hat{g}$ that implements the same outcome as some inalienable mechanism $(\bar{M}, \bar{g})$, the next proposition makes precise the converse of the previous proposition.

Proposition 4 *Consider a principal-agent problem Γ and the corresponding problem $\hat{\Gamma}$. If f is $\bar{M}$-implementable in Γ, then the social choice function $\hat{f} = (f, \bar{\phi}^1_{\bar{g}})$ is $\hat{g}$-implementable in $\hat{\Gamma}$, where $\bar{\phi}^1_{\bar{g}}$ is the first component of the response rule $\bar{\phi}_{\bar{g}}$ corresponding to the outcome function $\bar{g}$ satisfying (3).*

Proof of Proposition 4: Given f in Γ is $\bar{M}$-implementable, there is a $\bar{g}$ and an associated response rule $\bar{\phi}_{\bar{g}} = (\bar{\phi}^1_{\bar{g}}, \bar{\phi}^2_{\bar{g}})$ satisfying (3). Fixing functions $(\bar{g}, \bar{\phi}^1_{\bar{g}}, \bar{\phi}^2_{\bar{g}})$, consider the social choice function $\hat{f} = (f, \bar{\phi}^1_{\bar{g}})$ in $\hat{\Gamma}$ and the direct mechanism $\hat{g} = \hat{f}$. The propostion follows if $\hat{g}$ is incentive compatible, i.e. satisfies (2). To show this, fix a type $\theta \in \Theta$. It follows that $\hat{u}(\hat{g}(\theta), \theta) = \hat{u}(f(\theta), \bar{\phi}^1_{\bar{g}}(\theta)|\theta) = u(f(\theta), \theta)$, because $\hat{g} = \hat{f} = (f, \bar{\phi}_{\bar{g}})$ and $\bar{\phi}^1_{\bar{g}}(\theta) \in M(\theta)$. Hence, we have to show that $\hat{u}(\hat{g}(\theta')|\theta) = \hat{u}(f(\theta'), \bar{\phi}^1_{\bar{g}}(\theta')) \leq u(f(\theta), \theta)$ for any $\theta' \in \Theta$. Note first that while it holds $\bar{\phi}_{\bar{g}}(\theta') \in \bar{M}$, we can have $\bar{\phi}^1_{\bar{g}}(\theta') \notin M(\theta)$ or $\bar{\phi}^1_{\bar{g}}(\theta') \in M(\theta)$. First, suppose that $\bar{\phi}^1_{\bar{g}}(\theta') \notin M(\theta)$, it then follows $\hat{u}(\hat{g}(\theta')|\theta) = u(f(\theta'), \theta) - C \leq \max_{\tilde{x},\tilde{\theta}} u(\tilde{x}, \tilde{\theta}) - C = \min_{\tilde{x},\tilde{\theta}} u(\tilde{x}, \tilde{\theta}) \leq u(f(\theta), \theta)$. Next, suppose that $\bar{\phi}^1_{\bar{g}}(\theta') \in M(\theta)$, it then follows $\hat{u}(\hat{g}(\theta')|\theta) = u(f(\theta'), \theta) = u(\bar{g}(\bar{\phi}_{\bar{g}}(\theta')), \theta) \leq u(\bar{g}(\bar{\phi}_{\bar{g}}(\theta)), \theta)$, where the inequality follows because $\bar{\phi}_{\bar{g}}(\theta)$ maximizes $u(\bar{g}(m), \theta)$ over all $m \in \hat{M}(\theta) \cap \bar{M}$, which includes $\bar{\phi}_{\bar{g}}(\theta')$. Q.E.D.

Combining these two propositions with the previous two implies that Definition $\bar{1}$ gives a canonical representation of mechanisms in the sense that, in terms of implementable outcomes, there is no loss of generality in restricting attention to these mechanisms; any implementable outcome is implementable by some mechanism corresponding to Definition $\bar{1}$.

Example 1 revisited:
As an illustration to see how one can check the implementability of any social choice function in any principal-agent problem Γ and find the inalienable mechanism which implements it, reconsider the principal-agent problem $\Gamma_1 = (\Theta_1, X_1, M_1, u_1)$ of example 1 and the social choice function f_1. First construct the hypothetical principal-agent problem $\hat{\Gamma}_1 = (\Theta_1, X_1 \times \Theta_1, M_1, \hat{u}_1)$ where the hypothetical utility function $\hat{u}_1$ follows from its definition in (1):

$\hat{u}_1(x, \theta\|\theta_1)$	θ_1	θ_2	θ_3
x_1	10	10	0
x_2	15	15	5

$\hat{u}_1(x, \theta\|\theta_2)$	θ_1	θ_2	θ_3
x_1	-5	5	5
x_2	0	10	10

$\hat{u}_1(x, \theta\|\theta_3)$	θ_1	θ_2	θ_3
x_1	0	0	10
x_2	5	5	15

Next check whether there exists a social choice function $\hat{f}_1 = (f_1, \hat{\theta}_1)$ that is $\hat{g}$-implementable in $\hat{\Gamma}_1$. Given that the revelation principle holds in $\hat{\Gamma}$, this can be done as usual: find an incentive compatible direct mechanism $\hat{g}_1 = (\hat{g}^1_1, \hat{g}^2_1) : \Theta \to \hat{X}$ with $\hat{g}^1_1 = f_1$ and $\hat{g}^2_1 = \hat{\theta}_1$ which satisfies the familiar incentive compatible conditions (2). Using these incentive constraints one can verify that $\hat{g}_1(\theta_1) = (x_1, \theta_1)$, $\hat{g}_1(\theta_2) = \hat{g}_1(\theta_3) = (x_2, \theta_3)$ is such an incentive compatible direct mechanism. Hence, the conclusion follows that f_1 is indeed $\bar{M}$-implementable in Γ_1.

While this procedure confirms that f_1 is $\bar{M}$-implementable by the familiar means of checking incentive constraints of direct mechanisms, it does not yield the mechanism $(\bar{M}_1, \bar{g}_1)$ which actually implements f_1 in the principal-agent problem Γ_1. The constructive proof of Proposition 3 shows how to recover this mechanism from $\hat{g}_1$. Using that $\hat{g}_1 = (\hat{g}_1^1, \hat{g}_1^2) = (f_1, \hat{\theta}_1) = \hat{f}_1$, it follows

$$\bar{M}_1 = \{(\hat{\theta}_1(\theta_i), \theta_i) : \theta_i \in \Theta\} = \{(\hat{g}_1^2(\theta_i), \theta_i) : \theta_i \in \Theta\} = \{(\theta_1, \theta_1), (\theta_3, \theta_2), (\theta_3, \theta_3)\}.$$

This set yields the required $\bar{g}_1$ after linking it to the social choice function f_1 by setting $\bar{g}_1(\hat{\theta}(\theta), \theta) = \hat{f}_1^1(\theta) = f_1(\theta)$ for each $(\hat{\theta}(\theta), \theta) \in \bar{M}_1$. For Example 1, this yields $\bar{g}_1(\theta_1, \theta_1) = x_1$, $\bar{g}_1(\theta_3, \theta_2) = \bar{g}_1(\theta_3, \theta_3) = x_2$. □

6 Conclusion

This paper argues that, in their specific modeling of partially verifiable information, the seminal paper of Green and Laffont (1986) and subsequent literature implicitly restrict the notion of an implementable allocation. It, moreover, demonstrates a revelation principle in its full generality after expanding the notion of an implementable allocation without affecting the underlying information and verifiability structure. As usual, the obtained revelation principle allows a complete characterization of the set of implementable allocations by incentive constraints. In addition, it shows that any outcome associated with some (extended) implementable allocation is also implementable under the restricted interpretation of an allocation, provided that the mechanism designer is allowed to expand communication by adding unverifiable messages and to restrict communication by limiting the use of messages. These results lead to the conclusion that conceptual problems with the revelation principle are more related to a limited interpretation of an allocation and the implicit limitations on constructing these mechanisms rather than to the presence of verifiable information per se.

References

Ben-Porath, E. and B. Lipman (2012), "Implementation with partial provability," *Journal of Economic Theory*, 147, 1689-1724.

Bull, J. and J. Watson (2004), "Evidence disclosure and verifiability," *Journal of Economic Theory*, 118, 1-31.

Bull, J. and J. Watson (2007), "Hard evidence and mechanism design," *Games and Economic Behavior*, 58, 75-93.

Deneckere, R. and S. Severinov (2008), "Mechanism design with partial state verifiability," *Games and Economic Behavior*, 64, 487-513.

Forges, F. and F. Koessler (2005), "Communication equilibria with partially veriable types," *Journal of Mathematical Economics*, 41, 793-811.

Glazer, J. and A. Rubinstein (2004), "On optimal rules of persuasion," *Econometrica*, 72, 1715-1736.

Green, J., Laffont, J.-J. (1986), "Partially veriable information and mechanism design," *Review of Economic Studies*, 53, 447-456.

Hagenbach, J., F. Koessler, and E. Perez-Richet (2014), "Certiable Pre-Play Communication: Full Disclosure," *Econometrica*, 82, 1093-1131.

Kamenica, E. and M. Gentzkow (2011), "Bayesian Persuasion," *American Economic Review*, 101, 2590-2615.

Kartik, N. and O. Tercieux (2012), "Implementation with evidence," *Theoretical Economics*, 7, 323-355.

Lipman, B. and D. Seppi, 1995. "Robust inference in communication games with partial provability," *Journal of Economic Theory* , 66 (2), 370-405.

Postlewaite, A. and D. Schmeidler 1986. "Implementation in differential information economies," *Journal of Economic Theory*, 39, 14-33.

Singh, N. and D. Wittman (2001), "Implementation with partial verication," *Review of Economic Design*, 6, 63-84.