Information Games and Robust Trading Mechanisms

Gabriel Carroll, Stanford University

gdc@stanford.edu

April 5, 2019

Abstract

Agents about to engage in economic transactions may take costly actions to influence their own or others' information: costly signaling, information acquisition, hard evidence disclosure, and so forth. We study the problem of optimally designing a mechanism to be robust to all such activities, here termed *information games*. The designer cares about welfare, and explicitly takes the costs incurred in information games into account. We adopt a simple bilateral trade model as a case study. Any trading mechanism is evaluated by the expected welfare, net of information game costs, that it guarantees in the worst case across all possible games. Dominant-strategy mechanisms are natural candidates for the optimum, since there is never any incentive to manipulate information. We find that for some parameter values, a dominant-strategy mechanism is indeed optimal; for others, the optimum is a non-dominant-strategy mechanism, in which one party chooses which of two trading prices to offer.

Thanks to (in random order) Parag Pathak, Daron Acemoglu, Rohit Lamba, Laura Doval, Mohammad Akbarpour, Alex Wolitzky, Fuhito Kojima, Philipp Strack, Iván Werning, Takuro Yamashita, Juuso Toikka, and Glenn Ellison, as well as seminar participants at Michigan, Yale, Northwestern, CUHK, NUS, Chicago, Arizona State, Ohio State, and NYU, for helpful comments and discussions. Dan Walton provided valuable research assistance. The author is supported by a Sloan Research Fellowship, and also gratefully recognizes the hospitality of the Cowles Foundation at Yale during a visit there.

1 Introduction

Consider two agents about to engage in an economic transaction — say, sale of a good, at a price to be negotiated. Each agent may undertake various costly activities to influence her trading partner's information, or her own, to gain an advantage in bargaining: for example, costly learning about her partner's value for the good; efforts to prevent her partner from learning her own value; conversely, disclosing hard evidence of her own value; costly signaling as in Spence [26], where the meanings of the signals are determined within the equilibrium; and many other possibilities. Considerable research in game theory and information economics has gone into studying these kinds of activities, and their implications for equilibrium trading behavior, welfare, and optimal mechanism design.

In this paper, we take a broader perspective and consider a large class of such *information games* at once, which includes all of the above examples. An information game is any game (possibly dynamic) in which the agents take some actions, at some possible costs, and receive signals which may be informative about the values for the good. The benefits to receiving (and giving) such signals, and therefore the actions taken in the information game, naturally vary depending on what trading mechanism the agents will subsequently participate in. We therefore consider the following question: What mechanism would a planner adopt to maximize social welfare — where welfare takes into account both the outcome of the mechanism itself, and any costs spent on manipulating information beforehand? Rather than take a specific stand on the information game at hand, we envision a planner who is concerned about all such games. In the process, we also address the related question of what kind of information game makes it most difficult to achieve socially valuable trade. Note that the existence of an information game can potentially either hurt or help social welfare (it can be helpful by removing information frictions — in the extreme case where all agents become fully informed at no cost, the first-best becomes achievable); our conservative approach here focuses on hurtful cases.

In general, either with two agents or with more, one can consider agents learning about their own values, others' values, or both. Here we assume private values (agents already know their own preferences), so the learning is about others. This allows us to identify a natural focal class of mechanisms: *dominant-strategy* mechanisms, in which each agent is asked to report her preferences, and it is always in her best interest to be truthful, no matter what anyone else does. Indeed, in such a mechanism, no agent can ever have an incentive to spend any costs on affecting information (either her own or others') since it will not influence subsequent play of the mechanism.

This property has in fact been discussed in more applied contexts. In the school choice arena, for example, Pathak and Sönmez [22] present evidence of Boston parents strategically gathering information about others' choices. Moreover, Pathak and Sönmez [23] indicate that one explicit reason for England's shift away from the (non-dominant-strategy) Boston mechanism was that the latter "made the system unnecessarily complex ... [forcing] many parents to play an 'admissions game.'" This can be understood as a concern about the costs incurred by parents in strategizing. Milgrom's [20] review of market design similarly refers to participants' incentives to engage in espionage to learn about each other in non-dominant-strategy mechanisms; see also Li [17]. On the more theoretical side, Bikhchandani [4] also gives an example of how information acquisition can break non-dominant-strategy mechanisms. Our approach here allows us to ask rigorously whether this concern indeed justifies a focus on dominant-strategy mechanisms: If some other mechanism can generate better welfare — and can do so regardless of the information game at hand — then we reach a strong negative answer.

To study this question in detail, we need a specific application. As foreshadowed above, we adopt a version of the classic model of Myerson and Satterthwaite [21], where a buyer and seller meet to trade a single good. This is a natural choice because it is a canonical model with two important features:

- The designer's objective is welfare, so that it makes sense to be concerned with costs incurred in the information game.
- In the standard formulation of the problem, dominant-strategy mechanisms are suboptimal. (This is important; otherwise we would have no reason to consider non-dominant-strategy mechanisms.)

In addition, the quasi-linear utility makes it easy to combine allocative welfare and information game costs into a single objective.

In the usual treatment of this model, dominant-strategy mechanisms are simply posted prices (or randomizations over posted prices): the price is given, each party can accept or reject, and if both accept then trade occurs at that price [14]. This is not the case in our version because we consider discrete types. In Section 2, after presenting our basic bilateral trade model, we explicitly derive the optimal dominant-strategy mechanism, which involves probabilistic trade.

In spite of their considerable robustness, dominant-strategy mechanisms need not be optimal in our model. For some parameter values, the planner can guarantee a higher level of welfare — regardless of the information game — by using a *flexible-price* mechanism,

in which one agent can choose which of two prices to offer, and the other can accept or reject. Which price to offer depends on the offerer's belief about the receiver, and so there are incentives for the offerer to try to acquire information, and for the receiver to selectively reveal or hide information. Nonetheless, in equilibrium, we can bound the costs spent on these activities, and show that they can be outweighed by the efficiency gains from venturing outside the restrictive class of dominant-strategy mechanisms.

For what parameter values does this happen? For an intuition, note that the class of dominant-strategy mechanisms depends on the set of possible types of each agent, but not on their probabilities. Each such mechanism prescribes low probability of trade for some type profiles. When these type profiles have relatively high probability, this is when dominant-strategy mechanisms are too limiting, and flexible-price mechanisms can do better.

We would like to go beyond comparing two particular mechanisms, however, and actually optimize over all trading mechanisms. We can sharpen our question as follows: Each mechanism is evaluated by its robust guarantee across all information games, i.e. by the welfare that it generates in the worst case over information games. Here, welfare is defined in expectation (with respect to a given prior over agents' types) and, as stated above, accounts for both the material gains from trade in the mechanism and any costs (or benefits) incurred in the information game. Then we can ask, what exactly is the best such welfare guarantee, and what mechanism achieves it?

We focus on a very modest setting: the bilateral trade model with just two types of each agent. But for this setting, we answer our question completely (modulo some technicalities). For all parameter values, the optimal guarantee comes from either a dominant-strategy mechanism or a flexible-price mechanism, and we identify when each case occurs; see Figure 3 below.

The analysis also identifies the worst-case information game. This worst case is not one where agents pay to acquire information about each other, but rather one where they must pay to *prevent* information from being released. That is, the buyer is given a chance to pay some cost, and if she refuses, information becomes available that hurts the buyer's payoff in the trading mechanism; and the seller gets a chance to pay some cost, otherwise information becomes available that hurts the seller. Note that "hurting the buyer" does not necessarily mean that the seller becomes fully informed about the buyer's value; rather, the nature of the information that hurts the buyer is endogenous to the trading mechanism, and also may be different for the high-type buyer than the low type. Similarly for the seller.

This idea, that "informational extortion" serves as an adversarial information game, extends much more broadly than our specific bilateral trade model. Precursors to the informational extortion construction exist elsewhere in the game theory literature [15, 24], but making it precise in our setting requires much additional technical work, stemming primarily from the fact that any given trading mechanism may have multiple equilibria, and effective extortion depends on knowing which equilibrium is being played. All this is discussed in much more detail in Section 4.

This construction of informational extortion can be employed judiciously to give an upper bound on the welfare guarantee of any mechanism, and therefore of the best mechanism. On the other hand, we can also give a lower bound by simply analyzing the performance of particular mechanisms. The analysis outlined in Section 5 shows how the upper and lower bound coincide for all parameter values, thereby proving the main theorem.

This paper ties in thematically with several other recent robustness studies in mechanism design. Most closely related is the work by Brooks and Du [7], who consider a common-value auction model in which the information structure is unknown, and solve for the optimal auction under a maxmin criterion. (In the process, they also identify the worst-case information structure. See also [2].) Their objective is revenue, so the designer in their model cares only about the information that agents have when entering the mechanism, and not about costs they incur in getting to that information, as we do here.

Insofar as we assess the robustness of dominant-strategy mechanisms, this paper also connects with other studies on the foundations for such mechanisms. In particular, [9] considers that dominant-strategy mechanisms are robust to agents' beliefs about each other's types, and asks whether this robustness justifies focusing on such mechanisms (there looking at an auction context): If the designer evaluates a mechanism by worst-case revenue over all possible beliefs, is the optimal mechanism a dominant-strategy one? Similarly, [27] studies robustness to uncertainty about others' strategies, by assuming agents only play undominated strategies rather than equilibrium. Both of these papers, like ours, show that dominant-strategy mechanisms are indeed optimal for some parameters but not for others.[1] (Unlike them, the present paper identifies the optimum even when it is not a dominant-strategy mechanism.)

[1]Börgers [5] (see also [6]) observes that even if a dominant-strategy mechanism solves the maxmin problem, it may be weakly dominated in an appropriate sense. The criticism applies to our setting as well. However, the maxmin criterion is simple, and delivers (we hope) some relevant insights.

In one sense, the robustness criterion in the present paper is more demanding than those in [9, 27], and therefore more predisposed to select dominant-strategy mechanisms, since we consider not only uncertainty over agents' information but also care about costs incurred in attaining that information. However, our robustness criterion is also less demanding than theirs in that we restrict the space of possibilities by assuming equilibrium and common priors. Thus our criterion is not nested with those of [9] or [27]. Note that in our analysis here, if we did not impose any such "correct beliefs" assumptions, the planner would effectievly be restricted to dominant-strategy mechanisms: For any non-dominant-strategy mechanism, we could imagine an information game where an agent learns the other's type, and expects to pay nothing, but actually has to pay a million dollars; thus the mechanism leads to arbitrarily large welfare costs. We need to impose some modeling discipline to avoid such trivial conclusions, and we do this by assuming equilibrium throughout.

Finally, this paper naturally relates to the existing literature on mechanism design with endogenous information acquisition and other information games; although this literature has largely focused on agents learning about their *own* preferences (such as the classic studies [10, 12, 11]), rather than assuming, as we do here, that agents know their own preferences and are learning about others. Moreover, much of that literature has assumed very specific forms for the information game. An exception is Bergemann and Välimäki [3] who considered general information acquisition (about own preferences) and showed that the VCG mechanism is socially optimal, where the welfare criterion incorporates information acquisition costs.

2 The bilateral trade model

In this section, we lay out the basic bilateral trade model, without any information acquisition. It is a discrete-type version of the model of Myerson and Satterthwaite [21] with an exact budget-balance requirement.[2] (This two-type model was also studied by Matsuo [19].) In the following section, we will introduce information games and give the worst-case welfare criterion, and state the characterization of optimal mechanisms.

[2]We could instead consider weak budget balance, i.e. money can be thrown away but not created. This would require more variables, but would not change the substantive results.

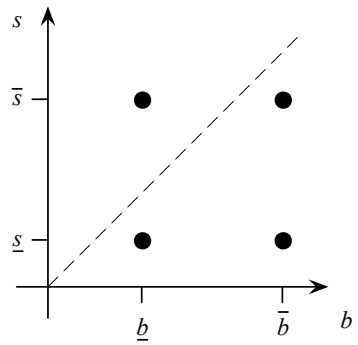

Figure 1: Possible values for each agent.

2.1 Definitions

There are two agents, a buyer (B) and a seller (S) of a good. The buyer's value for the good is b, and the seller's value (or cost of provision) is s. Each of these values has two possible realizations, $\underline{b}, \overline{b}$ for the buyer and $\underline{s}, \overline{s}$ for the seller (where $\underline{b} < \overline{b}$, $\underline{s} < \overline{s}$). Payoffs are quasi-linear, so if the interaction between the agents leads to a sale occurring with probability $q \in [0, 1]$, and the expected net payment from the buyer to seller is $t \in \mathbb{R}$, then the buyer's payoff is $qb - t$ and the seller's is $t - qs$; social welfare is the sum, $q(b - s)$.

The (common) prior is that the buyer's and seller's values are independently distributed, with probabilities $p_{\underline{b}}, p_{\overline{b}}$ for the buyer and $p_{\underline{s}}, p_{\overline{s}}$ for the seller (evidently $p_{\underline{b}} + p_{\overline{b}} = p_{\underline{s}} + p_{\overline{s}} = 1$). All these probabilities are assumed strictly positive. Each agent knows her own value at the time they interact.

The exogenous parameters of the model are the numbers $(\underline{b}, \overline{b}, \underline{s}, \overline{s}, p_{\overline{b}}, p_{\underline{s}})$. However, we will sometimes treat $\underline{b}, \overline{b}, \underline{s}, \overline{s}$ as fixed, and $p_{\overline{b}}, p_{\underline{s}}$ as variable parameters, in order to have a two-dimensional parameter space, which is conducive to drawing pictures.

We assume $\underline{s} < \underline{b} < \overline{s} < \overline{b}$. (It is straightforward to check that for any other ordering, first-best welfare can be achieved by a single posted price — a dominant-strategy mechanism — and so there is no reason for a planner to consider other mechanisms.) Thus the four possible pairs (b, s) are as shown in Figure 1. For the realization $(\underline{b}, \overline{s})$, it is socially optimal not to trade ($q = 0$), and for the other three realizations, it is optimal to trade ($q = 1$).

A planner designs the mechanism by which the agents will interact. Formally, an *indirect mechanism* — or a *mechanism* for short — is a quadruple (A_B, A_S, q, t), where:

- A_B, A_S are finite sets (specifying each agent's possible actions in the mechanism);

- $q : A_B \times A_S \to [0,1]$ and $t : A_B \times A_S \to \mathbb{R}$ are functions (representing probability of trade and net payment, as a function of the actions taken);
- there is a "non-participation" action $\emptyset \in A_B \cap A_S$, satisfying
  - $q(\emptyset, a_s) = 0$ and $t(\emptyset, a_s) \leq 0$ for all $a_s \in A_s$;
  - $q(a_b, \emptyset) = 0$ and $t(a_b, \emptyset) \geq 0$ for all $a_b \in A_b$.

The last requirement captures individual rationality — it ensures that each player can be guaranteed a payoff of at least zero by staying out.

We make two comments here on the modeling. First, we have required action sets to be finite; this assumption is made to avoid problems of equilibrium nonexistence, and it will also be imposed later when we introduce information games. Second, we have modeled mechanisms as effectively static. Later on we will use extensive-form games and extensive-form equilibrium refinements. It is therefore natural to consider mechanisms defined in extensive form. But by the end of the analysis it should be clear that this would not change our main results, so we simply keep the normal-form representation to save notation.

2.2 Dominant-strategy mechanisms

We begin the analysis by explicitly defining dominant-strategy mechanisms in our framework, and identifying the optimal such mechanism.

A *dominant-strategy mechanism* is one in which participants announce their values, and it is always optimal for them to do so truthfully. In our formalism, $A_B = \{\emptyset, \underline{b}, \overline{b}\}$ and $A_S = \{\emptyset, \underline{s}, \overline{s}\}$, but we can suppress the $\emptyset$ messages (assuming $q = t = 0$ whenever either a_b or a_s is $\emptyset$) and just represent the mechanism by functions $q : \{\underline{b}, \overline{b}\} \times \{\underline{s}, \overline{s}\} \to [0,1]$ and $t : \{\underline{b}, \overline{b}\} \times \{\underline{s}, \overline{s}\} \to \mathbb{R}$. These functions form a dominant-strategy mechanism if they satisfy the IC and IR constraints:

$$
\begin{aligned}
bq(b,s) - t(b,s) &\geq bq(b',s) - t(b',s) && \text{for each } b, b' \in \{\underline{b}, \overline{b}\} \text{ and } s \in \{\underline{s}, \overline{s}\};\\
bq(b,s) - t(b,s) &\geq 0 && \text{for each } b \in \{\underline{b}, \overline{b}\}, s \in \{\underline{s}, \overline{s}\};\\
t(b,s) - sq(b,s) &\geq t(b,s') - sq(b,s') && \text{for each } s, s' \in \{\underline{s}, \overline{s}\} \text{ and } b \in \{\underline{b}, \overline{b}\};\\
t(b,s) - sq(b,s) &\geq 0 && \text{for each } s \in \{\underline{s}, \overline{s}\}, b \in \{\underline{b}, \overline{b}\}.
\end{aligned}
$$

Expected welfare from the mechanism can be written as

$$W = p_{\underline{b}} p_{\underline{s}} (\underline{b} - \underline{s}) q(\underline{b}, \underline{s}) + p_{\underline{b}} p_{\overline{s}} (\underline{b} - \overline{s}) q(\underline{b}, \overline{s}) + p_{\overline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) q(\overline{b}, \underline{s}) + p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) q(\overline{b}, \overline{s}). \quad (2.1)$$

Noting that individual rationality forces $q(\underline{b}, \overline{s}) = t(\underline{b}, \overline{s}) = 0$, so that the second right-side term in (2.1) is zero, we can focus on the other three possible type profiles.

We can write $u_B(b, s) = bq(b, s) - t(b, s)$ and $u_S(b, s) = t(b, s) - sq(b, s)$ for the agents' payoffs at each type profile. Incentive-compatibility for the buyer (type $\overline{b}$ imitating $\underline{b}$) implies $u_B(\overline{b}, \underline{s}) \geq u_B(\underline{b}, \underline{s}) + (\overline{b} - \underline{b}) q(\underline{b}, \underline{s})$, and likewise incentive-compatibility for the seller (type $\underline{s}$ imitating $\overline{s}$) implies $u_S(\overline{b}, \underline{s}) \geq u_S(\overline{b}, \overline{s}) + (\overline{s} - \underline{s}) q(\overline{b}, \overline{s})$. It now is apparent that we cannot achieve first-best welfare in a dominant-strategy mechanism: First-best would mean $q(\underline{b}, \underline{s}) = q(\overline{b}, \underline{s}) = q(\overline{b}, \overline{s}) = 1$. But this would require

$$u_B(\overline{b}, \underline{s}) \geq u_B(\underline{b}, \underline{s}) + (\overline{b} - \underline{b}) q(\underline{b}, \underline{s}) \geq (\overline{b} - \underline{b})$$

and likewise

$$u_S(\overline{b}, \underline{s}) \geq (\overline{s} - \underline{s}),$$

so total welfare at profile $(\overline{b}, \underline{s})$ would satisfy

$$u_B(\overline{b}, \underline{s}) + u_S(\overline{b}, \underline{s}) \geq (\overline{b} - \underline{b}) + (\overline{s} - \underline{s}) = (\overline{b} - \underline{s}) + (\overline{s} - \underline{b}) > \overline{b} - \underline{s}$$

which is impossible.

We can extend this reasoning to identify the optimal (welfare-maximizing) dominant-strategy mechanism. The quantities $q(\underline{b}, \underline{s})$ and $q(\overline{b}, \overline{s})$ must satisfy a linear inequality that bounds them away from $(1, 1)$; there are two possible corner solutions depending on which one is equal to 1. This leads to the two (symmetrically equivalent) mechanisms described in Table 1. (The axes and labels have been oriented to be consistent with Figure 1.) Either may be optimal, depending on parameters.

Lemma 2.1. *(a) If* $p_{\underline{b}} p_{\underline{s}} \frac{\overline{b} - \underline{s}}{\overline{b} - \underline{b}} \geq p_{\overline{b}} p_{\overline{s}} \frac{\overline{b} - \overline{s}}{\overline{s} - \underline{s}}$, *then the mechanism shown on the left side of Table 1 is optimal among dominant-strategy mechanisms. The corresponding value of welfare is*

$$W_{DS} = p_{\underline{b}} p_{\underline{s}} (\underline{b} - \underline{s}) + p_{\overline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) + p_{\overline{b}} p_{\overline{s}} \frac{(\overline{b} - \overline{s})(\underline{b} - \underline{s})}{\overline{s} - \underline{s}}.$$

(b) If $p_{\underline{b}} p_{\underline{s}} \frac{\overline{b} - \underline{s}}{\overline{b} - \underline{b}} \leq p_{\overline{b}} p_{\overline{s}} \frac{\overline{b} - \overline{s}}{\overline{s} - \underline{s}}$, *then the mechanism shown on the right side of Table 1 is*

$\overline{s}$	$q:0$ $t:0$	$q:\frac{\overline{b}-\underline{s}}{\overline{s}-\underline{s}}$ $t:\frac{\overline{b}-\underline{s}}{\overline{s}-\underline{s}}\,\overline{s}$
$\underline{s}$	$q:1$ $t:\underline{b}$	$q:1$ $t:\underline{b}$
	$\underline{b}$	$\overline{b}$

$\overline{s}$	$q:0$ $t:0$	$q:1$ $t:\overline{s}$
$\underline{s}$	$q:\frac{\overline{b}-\overline{s}}{\overline{b}-\underline{b}}$ $t:\frac{\overline{b}-\overline{s}}{\overline{b}-\underline{b}}\,\underline{b}$	$q:1$ $t:\overline{s}$
	$\underline{b}$	$\overline{b}$

Table 1: Two possible forms for the optimal dominant-strategy mechanism.

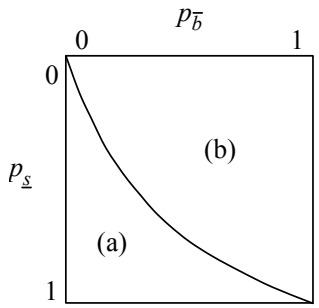

Figure 2: Regions of (probability) parameters where each case of Lemma 2.1 applies. (Here $(\underline{s},\underline{b},\overline{s},\overline{b}) = (1,2,3,5)$.)

optimal among dominant-strategy mechanisms. The corresponding value of welfare is

$$W_{DS} = p_{\underline{b}}p_{\underline{s}}\frac{(\overline{b}-\overline{s})(\underline{b}-\underline{s})}{\overline{b}-\underline{b}} + p_{\overline{b}}p_{\underline{s}}(\overline{b}-\underline{s}) + p_{\overline{b}}p_{\overline{s}}(\overline{b}-\overline{s}).$$

The full proof is in Appendix D.

The mechanism in case (a) can be interpreted as follows: The seller can choose to offer to sell the good at a price of $\underline{b}$; or she can offer a lottery in which, with probability $\frac{\overline{b}-\underline{s}}{\overline{s}-\underline{s}}$, the buyer buys at the higher price of $\overline{s}$ (and with remaining probability, no trade occurs). The buyer accepts if her value is at least the price offered (conditional on trade being realized). The mechanism in case (b) has a similar interpretation, with the buyer offering either to buy deterministically at price $\overline{s}$ or a lottery between buying at price $\underline{b}$ and nothing.

Each of the two mechanisms is optimal over a non-degenerate region of the $(p_{\overline{b}}, p_{\underline{s}})$ parameter space (for fixed $\underline{b},\overline{b},\underline{s},\overline{s}$). Figure 2 shows typical such regions.

2.3 No endogenous information

For contrast, we now briefly consider the Bayesian version of the problem, where the agents are presumed to have no information about each other's values (and no opportunities to acquire information). This is the version originally considered in [21].

In this case, we can again consider direct mechanisms, where each agent reports her value:[3] a mechanism is defined by functions $q : \{\underline{b}, \overline{b}\} \times \{\underline{s}, \overline{s}\} \to [0, 1]$ and $t : \{\underline{b}, \overline{b}\} \times \{\underline{s}, \overline{s}\} \to \mathbb{R}$ as before, but now they just need to satisfy the Bayesian versions of the IC and IR constraints:

$$\begin{aligned}
\sum_s p_s \left[bq(b,s) - t(b,s)\right] &\geq \sum_s p_s \left[bq(b',s) - t(b',s)\right] && \text{for each } b, b'; \\
\sum_s p_s \left[bq(b,s) - t(b,s)\right] &\geq 0 && \text{for each } b; \\
\sum_b p_b \left[t(b,s) - sq(b,s)\right] &\geq \sum_b p_b \left[t(b,s') - sq(b,s')\right] && \text{for each } s, s'; \\
\sum_b p_b \left[t(b,s) - sq(b,s)\right] &\geq 0 && \text{for each } s.
\end{aligned}$$

(Here the sums in the first two constraints are over $s \in \{\underline{s}, \overline{s}\}$; in the last two, over $b \in \{\underline{b}, \overline{b}\}$.)

Welfare for any such mechanism can be defined as in (2.1).

Our main observation is the following:

Proposition 2.2. *In the Bayesian problem, there exists a mechanism whose welfare is strictly higher than the best dominant-strategy mechanism.*

This can be seen by observing that the relevant incentive constraints are not binding at the optimal dominant-strategy mechanism, so we can improve on it by slightly increasing the probability of trade at the type profile that was originally constrained ($(\overline{b}, \overline{s})$ in case (a) of Lemma 2.1, $(\underline{b}, \underline{s})$ in case (b)). The formal proof is in Appendix D.

(With some further work one can identify the optimal Bayesian mechanism; see Matsuo [19]. For some parameters, even first-best welfare can be achieved. However, this is not important for our main goal.)

Proposition 2.2 will be useful as a benchmark, because our main result will show that, for some parameters, dominant-strategy mechanisms are optimal once information games

[3]The revelation principle says that the equilibrium outcome of any mechanism can be replicated by a direct mechanism, but we actually do not need this fact for present purposes.

are incorporated into the model. Comparing with Proposition 2.2 shows that we really do need information games to reach this conclusion.

One might ask whether there are easy arguments for optimality of dominant-strategy mechanisms by considering other canonical information structures. The answer is negative. For example, if both parties were fully-informed, then the first-best welfare could be attained in equilibrium. In fact if just one agent was fully-informed, this would already be sufficient (let the informed agent make a take-it-or-leave-it offer to the other). This subject will be addressed more thoroughly in Section 6.

3 Information games and welfare guarantees

We now complete the presentation of the model by describing how agents may acquire information. Then, we describe our central results.

3.1 Defining information games

Informally, we will define an information game to be any extensive-form game in which the players B and S move, possibly sequentially, and end up receiving some signals, which may be correlated with each other's values. We will require only that each agent has an "inaction" strategy available, of not actively spending anything to influence information.

To model these games, we will allow moves of nature to be correlated with the players' values (as we must, if nature is to send signals that are informative about the values). We will do this by explicitly modeling the probability space on which these moves are defined. Thus, for our purposes, define a *probability space* to be a tuple $\mathcal{P} = (\Omega, \pi, b, s)$ where

- Ω is a finite set;
- π is a full-support distribution over Ω;
- b, s are random variables on Ω, i.e. functions $b : \Omega \to \{\underline{b}, \overline{b}\}$, $s : \Omega \to \{\underline{s}, \overline{s}\}$, whose joint distribution follows the prior:

$$\pi(\{\omega \mid b(\omega) = \underline{b}, s(\omega) = \underline{s}\}) = p_{\underline{b}}p_{\underline{s}}, \qquad \pi(\{\omega \mid b(\omega) = \underline{b}, s(\omega) = \overline{s}\}) = p_{\underline{b}}p_{\overline{s}},$$
$$\pi(\{\omega \mid b(\omega) = \overline{b}, s(\omega) = \underline{s}\}) = p_{\overline{b}}p_{\underline{s}}, \qquad \pi(\{\omega \mid b(\omega) = \overline{b}, s(\omega) = \overline{s}\}) = p_{\overline{b}}p_{\overline{s}}.$$

(The use of the symbols b, s to denote both the random variables and specific realizations should not cause confusion in practice.)

Now we can be more specific: we define an *information game* to be a pair $\mathcal{I} = (\mathcal{P}, \mathcal{G})$, where $\mathcal{P}$ is a probability space, and $\mathcal{G}$ is a finite extensive-form game of perfect recall (as standardly defined, e.g. [13]) between players B and S, with the following modifications:

- in addition to the usual information partitions over nonterminal nodes, each player has an information partition over the terminal nodes, which also respects perfect recall;
- nature's moves at each relevant node are represented not by probability distributions but rather by mappings from possible states in Ω to successor nodes;
- at any two nodes in the same information set of player B, the random variable b has the same value, and similarly for S and s;
- there exists a strategy s_B for player B that guarantees a payoff of at least zero, i.e. $g_B(z) \geq 0$ for every terminal node z reachable under s_B (where g_B is B's payoff function), and similarly a strategy s_S for S that guarantees $g_S(z) \geq 0$.

This definition is, of course, still somewhat informal. The full formal definition is lengthy (as is unavoidable for extensive-form games) so we place it in Appendix A.

A few comments are in order on the approach we have adopted:

- Our definition imposes the assumption that players know their own values before participating in the information game. We could instead lift this assumption and assume only that players know their value by the time they participate in the mechanism. Doing so would allow a broader class of information games, but our main results would be unchanged.
- We have *not* assumed that players observe their realized payoffs in the information game (g_B, g_S) before the mechanism takes place. This could be added as a further restriction. Our main results would still hold, but the key ideas would be obscured: the proof of Theorem 3.3 below relies on a fairly involved construction, which would need to be further complicated to satisfy this extra restriction.
- We have allowed payoffs in the information game to be positive or negative. Costly information acquisition naturally suggests negative payoffs, but positive payoffs also

seem natural for some interactions (for example, one player selling verifiable information about his type to the other). Our Theorem 3.3 will require us to allow positive payoffs.

- If a player plays her inaction strategy, she can still passively receive information (besides her own value), as can her opponent. Thus our class of information games includes the possibility that information just arrives exogenously.

Now, given an information game $\mathcal{I}$ and a mechanism $\mathcal{M}$, they together form a *combined game*, in which the agents first play the information game and then (simultaneously) choose actions in the mechanism. Their payoffs are then

$$g_B(z) + bq(a_B, a_S) - t(a_B, a_S), \qquad g_S(z) + t(a_B, a_S) - sq(a_B, a_S)$$

where z is the terminal node reached in the information game, b, s the players' values there, and a_B, a_S the actions played in the mechanism. (Formal details are again in Appendix A.)

This defines the combined game as a standard, finite extensive-form game (with the one slightly unconventional feature that moves of nature are defined in terms of states, rather than probability distributions; it is easy to translate between the two). It describes the complete interaction of the buyer and the seller, starting from the ex-ante stage before their values are determined. Hence we can speak of strategies and equilibria in this game. Our basic solution concept will be sequential equilibrium (possibly in mixed strategies). We know that at least one such equilibrium exists.

Now, any mechanism $\mathcal{M}$ will be evaluated by its worst-case welfare, over all possible information games, where the welfare includes the payoffs incurred in the information game. That is, our criterion is the *welfare guarantee*

$$W(\mathcal{M}) = \inf_{\mathcal{I}} W(\mathcal{M}, \mathcal{I})$$

where $\mathcal{I}$ ranges over all information games, and $W(\mathcal{M}, \mathcal{I})$ is the total welfare (sum of the two agents' expected payoffs) in equilibrium of the combined game resulting from $\mathcal{I}$ and $\mathcal{M}$.

This definition is informal. In particular, a given combined game may have multiple equilibria, so which one should be used to define $W(\mathcal{M}, \mathcal{I})$? We will address this shortly. But first, we will introduce the rest of the concepts needed to describe our main results informally. We will then return to address technicalities, and formally define the welfare

criterion and state the results.

3.2 Dominant-strategy and flexible-price mechanisms

Let us consider now some specific mechanisms. For any dominant-strategy mechanism, we can see that its welfare guarantee $W(\mathcal{M})$ is simply the welfare defined in (2.1). Indeed, for any information game, each player has the option of earning a payoff at least zero in the information game (by inaction), and then the players simply play their dominant strategies in the mechanism. Thus, each player's equilibrium payoff in the combined game is at least her expected payoff in the mechanism alone. (It may be higher, if the information game allows positive payoffs. But this will not happen in the worst case.) Taking the best dominant-strategy mechanism, we can thus get a welfare guarantee of W_{DS}, as defined in Lemma 2.1.

We now consider an alternative: a mechanism in which one agent (say, the seller) can choose to offer trade at either of two prespecified prices, and the other agent can accept or reject. We will take the two prices to be $\underline{b}$ and $\overline{s}$; this will turn out to be optimal.

In our formalism (where mechanisms are represented as simultaneous-move games), we can represent the seller's actions as price offers $\underline{b}, \overline{s}$, and the buyer's action as the highest price she agrees to accept. Thus we define a *flexible-price mechanism (with seller offering)* as follows: $A_B = \{\emptyset, \underline{b}, \overline{s}\}$, $A_S = \{\emptyset, \underline{b}, \overline{s}\}$, and the functions q, t defining the mechanism are as in Table 2. (When $a_B = \emptyset$ or $a_S = \emptyset$, we take q, t to be zero.)[4]

$a_S = \overline{s}$	$q : 0$ $t : 0$	$q : 1$ $t : \overline{s}$
$a_S = \underline{b}$	$q : 1$ $t : \underline{b}$	$q : 1$ $t : \underline{b}$
	$a_B = \underline{b}$	$a_B = \overline{s}$

Table 2: Flexible-price mechanism (with seller offering).

We claim that such a mechanism guarantees expected welfare at least

$$W_{FP} = p_{\underline{s}}(\underline{b} - \underline{s}) + p_{\overline{b}}(\overline{b} - \overline{s}).$$

To see this, note that for any information game:

[4]This is a version of what Börgers and Smith [6] call a "downward flexible price mechanism."

- conditional on having value $\overline{b}$, the buyer gets an expected payoff of at least $\overline{b} - \overline{s}$ (in the combined game), since she always has the option of being inactive in the information game, and then accepting whichever price is offered;

- conditional on having value $\underline{s}$, the seller gets an expected payoff of at least $\underline{b} - \underline{s}$ (in the combined game), since she can always sit out of the information game and then offer price $\underline{b}$, which is always accepted;

- the buyer with value $\underline{b}$ and the seller with value $\overline{s}$ are assured payoffs at least zero.

Note that we could also have defined a flexible-price mechanism with the buyer offering (and the same choice of prices $\underline{b}, \overline{s}$); it would guarantee the same welfare level W_{FP}, by a symmetric argument.

Now we arrive at our first major observation: the flexible-price guarantee can be strictly higher than the dominant-strategy guarantee, for a non-negligible range of parameters. Indeed, we can identify explicitly the parameter region in which this happens. Define the following two thresholds for probabilities $p_{\overline{b}}, p_{\underline{s}}$:

$$p^*_{\overline{b}} = \frac{\underline{b} - \underline{s}}{\overline{b} - \underline{s}}, \qquad p^*_{\underline{s}} = \frac{\overline{b} - \overline{s}}{\overline{b} - \underline{s}}.$$

Note that both lie in the range $(0, 1)$.[5]

Proposition 3.1. *We have $W_{FP} \geq W_{DS}$ if and only if both $p_{\overline{b}} \leq p^*_{\overline{b}}$ and $p_{\underline{s}} \leq p^*_{\underline{s}}$ hold. Moreover, if $p_{\overline{b}} < p^*_{\overline{b}}$ and $p_{\underline{s}} < p^*_{\underline{s}}$, then $W_{FP} > W_{DS}$ strictly.*

The proof (by direct calculation) is in Appendix D.

For some intuition, temporarily relax the assumption that the buyer's and seller's values are independently distributed. Imagine instead that the prior is one of perfect correlation: the values are either $(\underline{b}, \underline{s})$ or $(\overline{b}, \overline{s})$. Then both parties are initially fully informed about each other's values. It is clear that the flexible-price mechanism can achieve all gains from trade — and with no incentive to spend to manipulate information, since there is already full information. On the other hand, dominant-strategy mechanisms perform strictly worse, since the probability of trade $q(\underline{b}, \underline{s})$ or $q(\overline{b}, \overline{s})$ in any such mechanism is bounded away from 1. In fact, even if the prior also places weight on $(\underline{b}, \overline{s})$ (where trade

[5]These thresholds have the following natural interpretation: $p^*_{\overline{b}}$ is the critical probability at which the seller with value $\underline{s}$, if she were able to offer any arbitrary price, would be indifferent between selling to both types of buyer (at price $\underline{b}$) or just to the high type (at price $\overline{b}$); similarly, $p^*_{\underline{s}}$ is the probability that makes a price-offering buyer of value $\overline{b}$ indifferent.

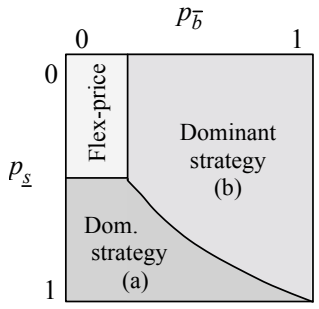

Figure 3: Parameter regions for the possible optimal mechanisms.

is inefficient), the flexible-price mechanism will still achieve all gains from trade without incentivizing spending on information — since the players are, in effect, fully informed *conditional* on gains from trade being available.

Now returning to full-support priors, a continuity argument suggests that flexible-price will continue to outperform dominant-strategy mechanisms as long as value realizations $(\underline{b}, \underline{s})$ and $(\overline{b}, \overline{s})$ are both very likely compared to the other state where trade is desirable, namely $(\overline{b}, \underline{s})$. Under our original assumption of independent values, this is true simply when $p_{\overline{b}}$ and $p_{\underline{s}}$ are both low — as reflected in the proposition.

3.3 The main result

Proposition 3.1 compares two guarantees from specific mechanisms. But what about the *optimal* mechanism, as measured by our worst-case guarantee criterion? This question leads to our main result: The optimum is either the dominant-strategy or the flexible-price mechanism (and Proposition 3.1 tells us which one, depending on parameters).

Theorem (informal).

(a) If $p_{\overline{b}} \geq p^*_{\overline{b}}$ or $p_{\underline{s}} \geq p^*_{\underline{s}}$, then the best welfare guarantee $W(\mathcal{M})$ of any mechanism $\mathcal{M}$ is W_{DS}, attained by the best dominant-strategy mechanism.

(b) If $p_{\overline{b}} < p^*_{\overline{b}}$ and $p_{\underline{s}} < p^*_{\underline{s}}$, then the best welfare guarantee $W(\mathcal{M})$ is W_{FP}, attained by a flexible-price mechanism (with either the seller or buyer offering).

The parameter regions for the various possibilities — two forms of dominant-strategy depending on the cases of Lemma 2.1, or flexible-price — are illustrated in Figure 3.

The statement above is informal because we still have not properly defined the welfare guarantee. There are two natural ways to define the equilibrium welfare of mechanism

$\mathcal{M}$ under information game $\mathcal{I}$, when multiple equilibria exist: we can consider the worst equilibrium, or the best equilibrium. Previous literature in robust mechanism design has used one or the other as is convenient (for example [2] uses the worst equilibrium; [1, 8] use the best). Here we will consider both, and will refer to them as $\underline{W}$ and $\overline{W}$ to distinguish.

Evaluating a mechanism by the worst equilibrium is naturally in the spirit of seeking robust guarantees. Note however that any mechanism always has an equilibrium in which each agent plays her non-participation strategy in the mechanism — it is a best reply to not participate if the opponent does the same — so we need some refinement to rule out this equilibrium in order to obtain a nontrivial welfare guarantee. We will impose here the assumption that agents play undominated actions in the mechanism. (Alternative refinements, such as trembling-hand perfect equilibrium, would give similar results.)

Formally: suppose we have a mechanism $\mathcal{M} = (A_B, A_S, q, t)$. For any mixed actions $\alpha_B \in \Delta(A_B)$, $\alpha_S \in \Delta(A_S)$, we can define $q(\alpha_B, \alpha_S)$ and $t(\alpha_B, \alpha_S)$ by taking expectations. Now for a buyer value $b \in \{\underline{b}, \overline{b}\}$, say that an action $a_B \in A_B$ is *weakly dominated given value* b if there exists a mixed action α'_B such that

$$bq(\alpha'_B, a_S) - t(\alpha'_B, a_S) \geq bq(a_B, a_S) - t(a_B, a_S) \qquad \text{for all } a_S \in A_S,$$

with strict inequality for some a_S. Similarly we define actions that are weakly dominated for each seller value $s \in \{\underline{s}, \overline{s}\}$. Notice that the set of weakly dominated actions for a given type of buyer (resp. seller) does not depend on any information that she may have about the seller (buyer).

Then, given $\mathcal{I}$, we consider sequential equilibria of the combined game in which each agent, at any information set where she takes an action in the mechanism, puts probability zero on any action that is weakly dominated given her value. For short, we call these *undominated sequential equilibria*. There always exists such an equilibrium (for example, any trembling-hand perfect equilibrium of the combined game).

We define $\underline{W}(\mathcal{M}, \mathcal{I})$ to be the lowest expected welfare, over all undominated sequential equilibria in the game formed from $\mathcal{M}$ and $\mathcal{I}$. We then define the corresponding welfare guarantee of a mechanism $\mathcal{M}$:

$$\underline{W}(\mathcal{M}) = \inf_{\mathcal{I}} \underline{W}(\mathcal{M}, \mathcal{I}). \tag{3.1}$$

The above operationalizes the worst-equilibrium criterion to evaluate a mechanism. An alternative is to instead use the *best* equilibrium. This also has methodological ad-

vantages. For example, part (a) of the main theorem shows that for some parameters, a designer can do no better than dominant-strategy mechanisms. This is clearly a stronger statement when a mechanism is evaluated by the best rather than the worst equilibrium. More importantly, by proving this statement under the best equilibrium, we make clear that the result is driven by the possibility of information acquisition, and not just by the equilibrium selection. Also, the best-equilibrium approach is in line with most of the mechanism design literature, where the implicit assumption is that the designer can instruct the agents on what strategies to play (as long as these strategies form an equilibrium).

However, our theorem using the best-equilibrium criterion will need one tweak to the model above: we must allow the information game to involve additional players besides the buyer and seller.[6] To see intuitively why this is helpful, imagine that the information game is being run by an adversary who wants to extract surplus from the traders. Because the mechanism may have multiple equilibria, the adversary wants to know which equilibrium is being played, and adapt the information game to that equilibrium; the additional players' role is to provide this information. (This will be discussed at greater length in the description of the worst-case information game in Section 4, and further in Section 6.)

Accordingly, we extend the definition of information games to allow extra players. These extra players need not have information partitions over the terminal nodes of the game (since they will not participate in the mechanism). We again require that each extra player have a strategy that guarantees her nonnegative payoff in the information game. The formal definition of information games, in Appendix A, makes all this precise. While incorporating additional players in the information game is somewhat inelegant, there is some precedent for it in the literature [18]. And arguably, it is defensible from the standpoint of realism: in reality, information exchange between two traders can involve strategic agents besides the traders themselves.

We can now define the equilibrium welfare of a mechanism $\mathcal{M}$, under information game $\mathcal{I}$, as the maximum sum of *all* players' expected payoffs (including the extra players) over all sequential equilibria of the combined game. Call this welfare level $\overline{W}(\mathcal{M}, \mathcal{I})$, and define the welfare guarantee of a mechanism by

$$\overline{W}(\mathcal{M}) = \inf_{\mathcal{I}} \overline{W}(\mathcal{M}, \mathcal{I}). \tag{3.2}$$

[6] This author does not currently know whether the theorem holds without allowing additional players; see Subsection 6.2.

Now we can properly state our main theorem, in two flavors depending on the equilibrium selection used. We note that we can also make the two versions of the theorem more comparable, if desired, by allowing additional players in both cases.

Theorem 3.2. *Define $\underline{W}(\mathcal{M})$ as in (3.1), where the infimum is over information games without additional players.*

(a) If $p_{\overline{b}} > p^*_{\overline{b}}$ or $p_{\underline{s}} > p^*_{\underline{s}}$, then the maximum value of $\underline{W}(\mathcal{M})$ over all mechanisms $\mathcal{M}$ is equal to W_{DS}.

(b) If $p_{\overline{b}} \leq p^*_{\overline{b}}$ and $p_{\underline{s}} \leq p^*_{\underline{s}}$, then the maximum value of $\underline{W}(\mathcal{M})$ over all mechanisms $\mathcal{M}$ is W_{FP}.

Moreover, if we instead define $\underline{W}(\mathcal{M})$ by taking the infimum over information games with additional players allowed, the same results hold.

Theorem 3.3. *Define $\overline{W}(\mathcal{M})$ as in (3.2), where the infimum is over information games with additional players allowed.*

(a) If $p_{\overline{b}} > p^*_{\overline{b}}$ or $p_{\underline{s}} > p^*_{\underline{s}}$, then the maximum value of $\overline{W}(\mathcal{M})$ over all mechanisms $\mathcal{M}$ is equal to W_{DS}. This maximum is attained by the dominant-strategy mechanism identified in Lemma 2.1.

(b) If $p_{\overline{b}} \leq p^*_{\overline{b}}$ and $p_{\underline{s}} \leq p^*_{\underline{s}}$, then the maximum value of $\overline{W}(\mathcal{M})$ is equal to W_{FP}. The maximum is attained by a flexible-price mechanism (with either agent offering).

The following sections describe the ideas of the proofs.

As one more technical note, observe that the worst-equilibrium version, Theorem 3.2, does *not* say that the dominant-strategy and second-price mechanisms attain the maximum guarantees. In fact, they do not, since we are looking at the worst equilibrium, and the undominated-action refinement is not enough to prevent some types from non-participation in these mechanisms. For example, in the dominant-strategy mechanism in case (a) of Lemma 2.1, the low-value buyer is completely indifferent between the truthful strategy $\underline{b}$ and the non-participation action $\emptyset$, so $\emptyset$ can be played in an undominated sequential equilibrium. However, this problem can be circumvented by adding extra actions so as to make non-participation become (weakly) dominated, as the full proof of the theorem shows. (Alternatively, we could keep the action set unchanged but perturb q and t so as to make non-participation become weakly dominated, and thereby approach — though not attain — the optimal guarantee.)

4 Informational extortion

In this section we present the key step in the proofs. This is the *extortion lemma*, a tool to upper-bound the welfare guarantee of any mechanism.

Here, first, is an informal description of how the bound arises. Imagine any information structure, describing the knowledge possessed by each agent when they enter the trading mechanism. For example, each may be fully informed about the other's value; or they may be completely uninformed; or each may have some independent noisy signal of the other's value. Clearly, the welfare guarantee of a mechanism $\mathcal{M}$ (measured either as $\underline{W}(\mathcal{M})$ or $\overline{W}(\mathcal{M})$; the difference does not matter for now) cannot exceed the expected welfare that arises if the players have the specified information and then play $\mathcal{M}$. And for any given information structure, it is routine (at least in principle) to compute the maximum possible welfare over all mechanisms. This number is thus a bound on the guarantee $W(\mathcal{M})$.

But we can find a potentially tighter bound as follows: Suppose $\mathcal{S}^1$ and $\mathcal{S}^2$ are two different information structures. Then, for any mechanism $\mathcal{M}$, its guarantee on the buyer's expected payoff is at most what she gets in $\mathcal{S}^1$, and its guarantee on the seller's expected payoff is at most what she gets in $\mathcal{S}^2$. To see this, let $\mathcal{S}^0$ be any "default" information structure, and now consider the following information game: Information is released according to structure $\mathcal{S}^1$, unless the buyer makes a payment to prevent it. If the buyer does make this payment, then information is released according to $\mathcal{S}^2$, unless the seller in turn makes a payment to prevent it. If both parties make their payments then $\mathcal{S}^0$ arises. The payments are calibrated to make each agent indifferent (or nearly indifferent) to paying. Then, indeed, the buyer's realized payoff in the combined game equals her payoff under $\mathcal{S}^1$ and the seller gets her payoff under $\mathcal{S}^2$. (This game should also make clear why we use the term "extortion." Simple versions of this construction have appeared in other contexts [15, 24], but additional work will be needed to implement the idea here.)

To apply this construction, we take $\mathcal{S}^1$ to be an information structure where the best possible expected payoff for the buyer, over all mechanisms, is relatively low, and take $\mathcal{S}^2$ to be an information structure that is likewise bad for the seller. This can give a bound on the sum of the two players' guarantees that is tighter than we could get by using any single information structure. Intuitively, we might expect that $\mathcal{S}^1$ involves the seller being fully informed and the buyer uninformed (so that any mechanism must give the seller some information rents), and $\mathcal{S}^2$ the reverse, but this will not always turn out to be the case.

We can further tighten this bound with some additional observations. First, we need not specify a priori which of $\mathcal{S}^1, \mathcal{S}^2$ is adversarial for the buyer and which is adversarial for the seller; this choice may be endogenous to the mechanism. A similar construction for the information game shows that each agent gets at most her payoff in whichever information structure is worst for her. Second and relatedly, we need not consider just two simultaneous information structures; we can equally well allow arbitrarily many (although the applications of these ideas in Section 5 will only require two). And third, the information structure that is worst for buyer type $\underline{b}$ may not be the same as for buyer type $\overline{b}$, and so we can further separate these types (and similarly for the seller).

In the rest of this section we develop these ideas formally. It should be apparent that the method and results can generalize straightforwardly to arbitrarily many types of buyer and seller, and indeed, that they can be applied not just to the bilateral trade setting but to many other mechanism design problems. The ideas may therefore be of independent interest. However, to avoid introducing yet more notation, we will state them here in the specific context of the two-type bilateral trade model.

4.1 Information structures and the extortion lemma

We define an *information structure* to be $\mathcal{S} = (\Omega, \pi, b, s, H_B, H_S, \eta_B, \eta_S)$ where

- (Ω, π, b, s) is a probability space;
- H_B and H_S are finite sets, representing the possible *signals* of the buyer and seller;
- $\eta_B : \Omega \to H_B$ and $\eta_S : \Omega \to H_S$ are surjective functions, such that if $\eta_B(\omega) = \eta_B(\omega')$ then $b(\omega) = b(\omega')$, and if $\eta_S(\omega) = \eta_S(\omega')$ then $s(\omega) = s(\omega')$.

We may also use η_B, η_S as variables for representative elements of H_B, H_S (signals). Since π is required to have full support, and η_B, η_S are surjective, every signal has positive probability. Note that the last requirement effectively says that each agent knows her own value, and consequently we may write $b(\eta_B)$ or $s(\eta_S)$ with clear meaning.

It will sometimes be useful to abbreviate the probability of a given signal, or profile of signals, as

$$\pi(\eta_B^*) = \sum_{\omega:\eta_B(\omega)=\eta_B^*} \pi(\omega), \qquad \pi(\eta_S^*) = \sum_{\omega:\eta_S(\omega)=\eta_S^*} \pi(\omega), \qquad \pi(\eta_B^*, \eta_S^*) = \sum_{\substack{\omega:\eta_B(\omega)=\eta_B^*,\\ \eta_S(\omega)=\eta_S^*}} \pi(\omega). \tag{4.1}$$

We may also notate the conditional probabilites as

$$\pi(\eta_B^*|\eta_S^*) = \frac{\pi(\eta_B^*, \eta_S^*)}{\pi(\eta_S^*)}, \qquad \pi(\eta_S^*|\eta_B^*) = \frac{\pi(\eta_B^*, \eta_S^*)}{\pi(\eta_B^*)}$$

(unambiguous since the denominators are positive).

Fix an information structure $\mathcal{S}$, and suppose the players have the information specified and then play some mechanism. The revelation principle tells us that the (expected) outcome can be described by some direct mechanism, where the players simply report their signals. This motivates us to define a *direct mechanism on* $\mathcal{S}$ as a pair of functions (q, t), with $q : H_B \times H_S \to [0, 1]$ and $t : H_B \times H_S \to \mathbb{R}$, satisfying the IC and IR constraints

$$\sum_{\eta_S} \pi(\eta_S|\eta_B^*) \left[b(\eta_B^*)q(\eta_B^*, \eta_S) - t(\eta_B^*, \eta_S)\right] \geq \sum_{\eta_S} \pi(\eta_S|\eta_B^*) \left[b(\eta_B^*)q(\eta_B', \eta_S) - t(\eta_B', \eta_S)\right]$$
$$\text{for each } \eta_B^*, \eta_B' \in H_B;$$
$$\sum_{\eta_S} \pi(\eta_S|\eta_B^*) \left[b(\eta_B^*)q(\eta_B^*, \eta_S) - t(\eta_B^*, \eta_S)\right] \geq 0 \quad \text{for each } \eta_B^* \in H_B;$$
$$\sum_{\eta_B} \pi(\eta_B|\eta_S^*) \left[t(\eta_B, \eta_S^*) - s(\eta_S^*)q(\eta_B, \eta_S^*)\right] \geq \sum_{\eta_B} \pi(\eta_B|\eta_S^*) \left[t(\eta_B, \eta_S') - s(\eta_S^*)q(\eta_B, \eta_S')\right]$$
$$\text{for each } \eta_S^*, \eta_S' \in H_S;$$
$$\sum_{\eta_B} \pi(\eta_B|\eta_S^*) \left[t(\eta_B, \eta_S^*) - s(\eta_S^*)q(\eta_B, \eta_S^*)\right] \geq 0 \quad \text{for each } \eta_S^* \in H_S.$$

We may denote a direct mechanism (q, t) by the single variable $\mathcal{M}$. Although the notations $\mathcal{M}, q, t$ are also used for indirect mechanisms, no confusion should result.

We can define the expected payoff of each type of buyer and seller in a direct mechanism $\mathcal{M} = (q, t)$:

$$u_{\underline{b}}(\mathcal{M}) = \frac{1}{p_{\underline{b}}} \times \sum_{(\eta_B, \eta_S): b(\eta_B) = \underline{b}} \pi(\eta_B, \eta_S) \left[\underline{b}q(\eta_B, \eta_S) - t(\eta_B, \eta_S)\right];$$
$$u_{\overline{b}}(\mathcal{M}) = \frac{1}{p_{\overline{b}}} \times \sum_{(\eta_B, \eta_S): b(\eta_B) = \overline{b}} \pi(\eta_B, \eta_S) \left[\overline{b}q(\eta_B, \eta_S) - t(\eta_B, \eta_S)\right];$$
$$u_{\underline{s}}(\mathcal{M}) = \frac{1}{p_{\underline{s}}} \times \sum_{(\eta_B, \eta_S): s(\eta_S) = \underline{s}} \pi(\eta_B, \eta_S) \left[t(\eta_B, \eta_S) - \underline{s}q(\eta_B, \eta_S)\right];$$
$$u_{\overline{s}}(\mathcal{M}) = \frac{1}{p_{\overline{s}}} \times \sum_{(\eta_B, \eta_S): s(\eta_S) = \overline{s}} \pi(\eta_B, \eta_S) \left[t(\eta_B, \eta_S) - \overline{s}q(\eta_B, \eta_S)\right].$$

Our extortion lemma will involve multiple information structures. Thus, we will refer to a finite list of the form $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ as a *list of information structures*, and given such a list, we can refer to a *list of direct mechanisms* on it, $\mathcal{L} = (\mathcal{M}^1, \ldots, \mathcal{M}^K)$, where each $\mathcal{M}^k$ is a direct mechanism on $\mathcal{S}^k$ $(k = 1, \ldots, K)$.

Given such a list of information structures and corresponding list of direct mechanisms, we can define the *minimum utility* of each type of agent as her worst payoff over all the mechanisms in the list:

$$\begin{aligned} u_{\underline{b}}(\mathcal{L}) &= \min_k u_{\underline{b}}(\mathcal{M}^k); \\ u_{\overline{b}}(\mathcal{L}) &= \min_k u_{\overline{b}}(\mathcal{M}^k); \\ u_{\underline{s}}(\mathcal{L}) &= \min_k u_{\underline{s}}(\mathcal{M}^k); \\ u_{\overline{s}}(\mathcal{L}) &= \min_k u_{\overline{s}}(\mathcal{M}^k). \end{aligned}$$

And finally, we can define the *total minimum utility* of the list as the (probability-weighted) sum of the minimum utility of each type of each agent:

$$TMU(\mathcal{L}) = p_{\underline{b}} u_{\underline{b}}(\mathcal{L}) + p_{\overline{b}} u_{\overline{b}}(\mathcal{L}) + p_{\underline{s}} u_{\underline{s}}(\mathcal{L}) + p_{\overline{s}} u_{\overline{s}}(\mathcal{L}). \tag{4.2}$$

Now we come to the first complete statement of the extortion lemma. It says that, for any given list of information structures, the guarantee of any mechanism is bounded by the best possible TMU of any direct mechanism list. The lemma again comes in two flavors, depending which equilibrium selection criterion we use.

Lemma 4.1. *Let $\mathcal{M}$ be any mechanism. Define $\underline{W}(\mathcal{M})$ as in (3.1), where the infimum is taken over information games without additional players. Let $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ be any list of information structures. Then*

$$\underline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}),$$

where the max is over all lists of direct mechanisms $\mathcal{L} = (\mathcal{M}^1, \ldots, \mathcal{M}^k)$ for the given information structures.

(If we instead define $\underline{W}$ by taking the infimum over information games with additional players allowed, the same result holds a fortiori.)

Lemma 4.2. *Let $\mathcal{M}$ be any mechanism. Define $\overline{W}(\mathcal{M})$ as in (3.2), where the infimum is over information games with additional players allowed. Let $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ be any list of*

information structures. Then

$$\overline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}),$$

where the max is over all lists of direct mechanisms for the given information structures.

(It is straightforward to check that the max on the right-hand side is attained.)

In fact, our proofs of the main theorems will use a version of the extortion lemma with a slight technical strengthening.[7] Suppose that the information structures $\mathcal{S}^k$ overlap, in the sense that a portion of one information structure is isomorphic to a portion of another information structure. We can then impose that the corresponding direct mechanisms behave in the same way across both information structures. This is a restriction on the possible lists of direct mechanisms, and so it (potentially) makes the bound given by the extortion lemma tighter.

Specifically: Given a list of information structures $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$, we use notation $(\Omega^k, \pi^k, b^k, s^k, H_B^k, H_S^k, \eta_B^k, \eta_S^k)$ to refer to the components of information structure k. Suppose we have a list of information structures, in which the same buyer signal may appear in more than one information structure, and similarly for seller signals. We say that the list is an *overlapping list of information structures* if it satisfies the following properties:

- For each k and k', if $\eta_B \in H_B^k \cap H_B^{k'}$, then $b^k(\eta_B) = b^{k'}(\eta_B)$, and for each $\eta_S \in H_S^k \cup H_S^{k'}$,
$$\pi^k(\eta_S|\eta_B) = \pi^{k'}(\eta_S|\eta_B).$$
- For each k and k', if $\eta_S \in H_S^k \cap H_S^{k'}$, then $s^k(\eta_S) = s^{k'}(\eta_S)$, and for each $\eta_B \in H_B^k \cup H_B^{k'}$,
$$\pi^k(\eta_B|\eta_S) = \pi^{k'}(\eta_B|\eta_S).$$

Here the probabilities $\pi^k(\eta_B)$, etc. are defined as in (4.1), and we take $\pi^k(\eta_B, \eta_S) = 0$ if $\eta_B \notin H_B^k$ or $\eta_S \notin H_S^k$. Thus, the requirement says that any given signal should convey the same information, both about values and about the other player's signal, in each information structure where it can arise.

In particular, we can write the conditional probabilities $\pi(\eta_S|\eta_B)$, $\pi(\eta_B|\eta_S)$ without needing to specify k. We also can write $b(\eta_B), s(\eta_S)$ for the values associated with these signals, again without ambiguity.

[7] This author does not know whether the unstrengthened version of the lemma is already enough to prove the main theorem. However, for the specific choice of information structures used in the proof here, the strengthening is needed; see Appendix E.3.

If $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ is an overlapping list of information structures, then we say that $\mathcal{L} = (\mathcal{M}^1, \ldots, \mathcal{M}^K)$ is an *overlapping list of direct mechanisms* on it if each $\mathcal{M}^k = (q^k, t^k)$ is a direct mechanism on $\mathcal{S}^k$, and for every signal pair (η_B, η_S) that appears in two different information structures, $\eta_B \in H_B^k \cap H_B^{k'}$ and $\eta_S \in H_S^k \cap H_S^{k'}$, we have

$$q^k(\eta_B, \eta_S) = q^{k'}(\eta_B, \eta_S) \qquad \text{and} \qquad t^k(\eta_B, \eta_S) = t^{k'}(\eta_B, \eta_S). \tag{4.3}$$

Thus, the mechanisms should respect the overlaps across information structures. We then define the minimum utility of each type, and the total minimum utility, as before.

We can now state the strengthened extortion lemma:

Lemma 4.3. *Let $\mathcal{M}$ be any mechanism. Define $\underline{W}(\mathcal{M})$ as in (3.1), where the infimum is taken over information games without additional players. Let $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ be an overlapping list of information structures. Then*

$$\underline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}),$$

where the max is over overlapping lists of direct mechanisms for the given information structures.

(If we instead define $\underline{W}$ by taking the infimum over information games with additional players allowed, the same result holds a fortiori.)

Lemma 4.4. *Let $\mathcal{M}$ be any mechanism. Define $\overline{W}(\mathcal{M})$ as in (3.2), where the infimum is over information games with additional players allowed. Let $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ be an overlapping list of information structures. Then*

$$\overline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}),$$

where the max is over overlapping lists of direct mechanisms for the given information structures.

Ahead, we give an outline of the proofs. For intuition it suffices to focus on the non-overlapping versions of the extortion lemma. The formal proofs are in Appendix B, and they cover the overlapping versions, Lemmas 4.3 and 4.4. (Lemmas 4.1 and 4.2 immediately follow.)

4.2 Proof sketches

Worst-equilibrium criterion. The argument for the extortion lemma with the worst-

equilibrium criterion is largely as outlined at the start of this section. Take the mechanism $\mathcal{M}$, and the list of information structures $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$, and also let $\mathcal{S}^0$ be an arbitrary "default" information structure. For each $k = 0, 1, \ldots, K$, imagine a game in which the players exogenously receive signals η_B and η_S according to $\mathcal{S}^k$, and then play the mechanism. Fix an equilibrium of this game, in undominated actions. The (expected) equilibrium outcome defines a direct mechanism $\mathcal{M}^k$ on $\mathcal{S}^k$. This gives us a list of direct mechanisms $\mathcal{L}$. It suffices to show that the guarantee of $\mathcal{M}$ is at most the TMU of this list.

Define $\underline{k}(\underline{b})$ as the value of $k \in \{1, \ldots, K\}$ for which $u_{\underline{b}}(\mathcal{M}^k)$ is lowest, and define $\Delta_{\underline{b}} = u_{\underline{b}}(\mathcal{M}^0) - u_{\underline{b}}(\mathcal{M}^{\underline{k}(\underline{b})})$. This latter quantity is type $\underline{b}$'s willingness to pay to face information structure $\mathcal{S}^0$ rather than $\mathcal{S}^{\underline{k}(\underline{b})}$ (taking as given the play of the mechanism in each information structure). Similarly we can define $\underline{k}(\overline{b}), \underline{k}(\underline{s}), \underline{k}(\overline{s})$ and $\Delta_{\overline{b}}, \Delta_{\underline{s}}, \Delta_{\overline{s}}$.

Now, consider the information game structured as follows:

- First, the buyer has the chance to pay Δ_b (that is, to pay $\Delta_{\underline{b}}$ if the buyer's value is $\underline{b}$, and $\Delta_{\overline{b}}$ if $\overline{b}$). The buyer can accept or reject this "extortion" opportunity.
- Then, if the buyer has accepted, the seller has the chance to pay Δ_s. The seller can accept or reject.
- An information structure is chosen as follows. If both players accepted, then $k = 0$. If the buyer rejected, then $k = \underline{k}(b)$. If the buyer accepted but the seller rejected, then $k = \underline{k}(s)$.
- Both parties receive signals (η_B, η_S) according to the chosen information structure $\mathcal{S}^k$. Payoffs in the information game are: $-\Delta_b$ for the buyer if she accepted, and 0 if she rejected; similarly for the seller.

This certainly describes an information game; note that each player does indeed have an inaction strategy (rejecting the extortion offer). It is natural to then argue that the players can play as follows: each player accepts the extortion, and when information structure $\mathcal{S}^k$ is realized, they play the equilibrium corresponding to $\mathcal{M}^k$. Note that the payments are calibrated to make each player indifferent between accepting or rejecting extortion, so that they are willing to accept in equilibrium. In particular, this means that type $\underline{b}$'s expected payoff in equilibrium is $u_{\underline{b}}(\mathcal{M}^{\underline{k}(\underline{b})}) = u_{\underline{b}}(\mathcal{L})$. Similarly for types $\overline{b}, \underline{s}, \overline{s}$. Therefore total welfare in the combined game is $TMU(\mathcal{L})$.

Actually, this construction does not quite work. There are two issues. First, suppose $\underline{k}(\underline{s}) \neq \underline{k}(\overline{s})$, and consider an out-of-equilibrium information set (in the combined game) where the buyer observes a signal that exists in information structure $\mathcal{S}^{\underline{k}(\underline{s})}$, but not in $\mathcal{S}^{\underline{k}(\overline{s})}$. She then knows that the seller is type $\underline{s}$ (and unexpectedly rejected extortion). Consequently, receiving this signal gives the buyer additional information, besides that contained in $\mathcal{S}^{\underline{k}(\underline{s})}$. We then cannot expect her to play according to the original equilibrium of $\mathcal{S}^{\underline{k}(\underline{s})}$.

To avoid this problem, we must add a trembling stage to the information game: After the decisions have been made to accept or reject extortion, with exogenous probability ϵ, the extortion decisions are ignored, and instead an information structure $k \in \{0, 1, \ldots, K\}$ is drawn at random. (In this case, any previously accepted payments are not made.) These changes ensure that every information structure arises with positive probability on path, and so when either player finds that an information structure $\mathcal{S}^k$ has "unexpectedly" arisen, she believes that it arose due to the tremble in the information game, and draws no further inferences about the other player.

The second, and related, issue is that with the timing above, in the (off-equilibrium) situation where the seller never receives an extortion offer, she knows the buyer has rejected extortion. Then, she can infer the buyer's value, based on whether she receives a signal from $\mathcal{S}^{\underline{k}(\underline{b})}$ or $\mathcal{S}^{\underline{k}(\overline{b})}$. Thus the buyer's rejection gives the seller more information than intended. Note that trembles do not fix this issue (they make this inference imperfect but still informative). Instead, we fix it by making the buyer's and seller's decisions *simultaneous* — but then using the seller's decision only if the buyer accepts. This way, the seller doesn't know if the buyer has rejected.

A full description of the information game used is in Appendix B.1. It is also described schematically in Figure 4 below. Note that the figure does not show the full game tree (which would involve many more nodes, due to moves of nature, first in assigning values and then drawing signal realizations).

Best-equilibrium criterion. When we move to use the best-equilibrium selection criterion, further complications arise and we outline them here, leaving a full description of the information game to Appendix B.3. In this case, we need an information game that ensures that *every* equilibrium leads to welfare at most $\max_{\mathcal{L}} TMU(\mathcal{L})$ (or close to it).

The construction used above is not sufficient. To see why, note that this construction depended on a particular prescription for equilibrium play of $\mathcal{M}$ under each information structure $\mathcal{S}^k$, in order to define the quantities Δ_b, Δ_s. However, if this information game arises, the players will not necessarily play the specified equilibrium of $\mathcal{M}$; they might

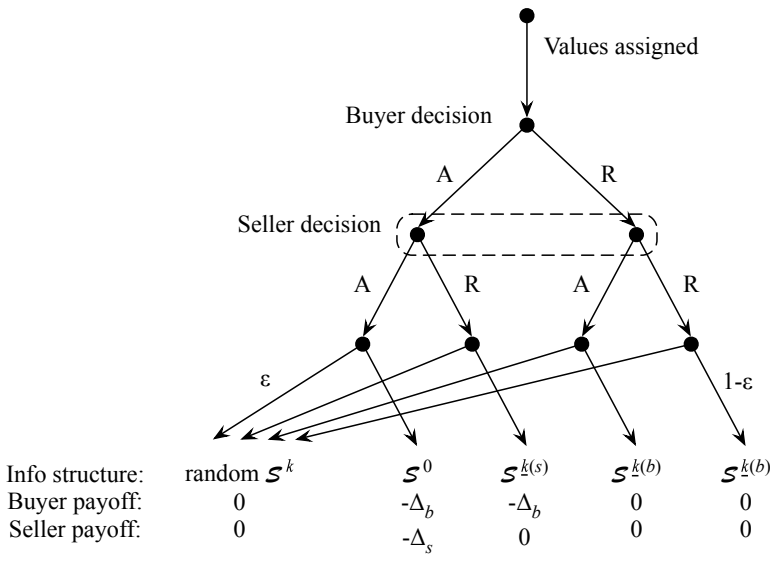

Figure 4: Adversarial information game, under the worst-equilibrium selection.

follow a different equilibrium. Then, Δ_b no longer represents the difference in B's equilibrium payoff between $\mathcal{S}^0$ and $\mathcal{S}^{\underline{k}(b)}$ (and likewise for S), so the information game fails to push each player down to her payoff in the worst information structure.

Instead, we need an information game that endogenously elicits the quantities Δ_b and Δ_s (and the values $\underline{k}(b), \underline{k}(s)$), so that they correspond to whichever actual equilibrium is being played. This is why we introduce additional players. Specifically, we include two additional players, "informants" whom we call I_B and I_S. I_B first reports a quadruple $(\underline{k}(\underline{b}), \underline{k}(\overline{b}), \Delta_{\underline{b}}, \Delta_{\overline{b}})$, describing each buyer type's least-preferred information structure and willingness-to-pay to avoid it; and similarly for I_S and the seller. Then the buyer and seller are simultaneously given extortion offers as before: the buyer is extorted via the (reported) $(\underline{k}(\underline{b}), \Delta_{\underline{b}})$ if her value is $\underline{b}$, and $(\underline{k}(\overline{b}), \Delta_{\overline{b}})$ if her value is $\overline{b}$; similarly for the seller. The realized information structure is then determined according to the two agents' responses to the extortion offers, with ϵ chance of trembling to a random information structure, as before. One additional change needed is that, instead of processing the buyer's acceptance or rejection before the seller's, we must process them in a random order, in order to ensure that both players have positive probability of influencing the resulting information structure in every equilibrium (this ensures that their response to any extortion offer really does reflect willingness to pay). Finally, the informants are incentivized to make reports that correspond to the traders' actual willingness to pay, by giving them small payments that are increasing in the reported amounts Δ_b (respectively

Δ_s), but imposing large fines if the extortion offers are rejected.

An additional subtlety arising in the proof is that (say) "the buyer's payoff under $\mathcal{S}^k$" can mean two things: it can mean the expected payoff in equilibrium conditional on reaching $\mathcal{S}^k$, or it can mean the buyer's payoff after a *deviation* where she rejects extortion and thereby lands in $\mathcal{S}^k$. The mechanism $\mathcal{M}^k$ is constructed from the former, but the extortion procedure reflects the latter; so the proof can succeed only if these two measures of the buyer's payoff in $\mathcal{S}^k$ are (approximately) equal. Now, if the seller were to reject extortion with non-negligible probability in equilibrium, then the two measures might be different, because these two different avenues for reaching $\mathcal{S}^k$ might convey different information to the buyer about whether the seller had rejected, and this might in turn be correlated with the seller's future play. However, by making the fines on rejected informants sufficiently large, we ensure that in every equilibrium, the buyer and seller both have very small probability of rejecting extortion (otherwise some informant would have negative expected payoff and so would rather deviate to inaction). Then this whole difficulty is avoided.

The details of the game construction, and the proof that it ensures the desired upper bound on welfare, are in Appendix B. Subsection 6.2 below contains more discussion about the role of informants in the construction.

5 Characterizing the optimal mechanism

The proof of our characterization of the optimum, Theorems 3.2 and 3.3, now proceeds by applying the extortion lemma. Note that for any given list of information structures, the problem of maximizing $TMU(\mathcal{L})$ over all direct mechanism lists $\mathcal{L}$ is a linear program: The variables are the direct mechanism parameters $q^k(\eta_B, \eta_S)$ and $t^k(\eta_B, \eta_S)$, and the values $u_{\underline{b}}, u_{\overline{b}}, u_{\underline{s}}, u_{\overline{s}}$; the objective is to maximize $p_{\underline{b}}u_{\underline{b}} + p_{\overline{b}}u_{\overline{b}} + p_{\underline{s}}u_{\underline{s}} + p_{\overline{s}}u_{\overline{s}}$, subject to the IC and IR constraints for each mechanism, *and* the constraints that the utility of type $\underline{b}$ in each mechanism $\mathcal{M}^k$ should be at least $u_{\underline{b}}$, and similarly for $\overline{b}, \underline{s}, \overline{s}$. (For overlapping lists, we have the additional constraints (4.3).) Hence, in principle, solving this maximization problem is mechanical.

Thus, for an appropriately chosen list of information structures, we can calculate the maximum value of $TMU(\mathcal{L})$ over all $\mathcal{L}$. According to the extortion lemma, this is an upper bound on the welfare guarantee of any possible mechanism $\mathcal{M}$. Showing that this bound is attained (by an appropriate version of the dominant-strategy or flexible-price mechanism) then completes the characterization.

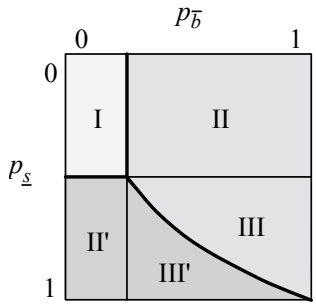

Figure 5: Parameter regions for the proof of the main characterization.

In order to get a tight bound, we just have to find the right list of information structures. For this, we break into cases depending on the parameters. Figure 5 shows the parameter space carved into five regions. We describe here the information structures used in regions $\mathrm{I}, \mathrm{II}, \mathrm{III}$. (Regions II' and III' are symmetric.) The upper bound for each of these regions is stated as a lemma.

The (mechanical) proofs of these upper bounds are in Appendix C. Then, the formal proofs of Theorems 3.2 and 3.3 just consist of tying together the pieces; these proofs are in Appendix D.

5.1 Region I

In this region, $p_{\overline{b}} \leq p^*_{\overline{b}}$ and $p_{\underline{s}} \leq p^*_{\underline{s}}$.

Here we can start from a straightforward guess at the worst information structures: the worst information structure for the buyer is one where the seller is fully informed (and the buyer knows only her value); the worst information structure for the seller is the reverse.

Morally this is right, but it will be helpful for us to make two adjustments:

- First, we separate out the state in which the values are $(\underline{b}, \overline{s})$ so that trade is not valuable; that is, we assume both agents find out whether this value profile has occurred. If it has, then the IR constraints imply zero trade and zero payment. This leaves us fewer remaining states to worry about.

  Thus, the resulting information structures are as shown in Table 3. Here, in each of the two information structures, there are four states, corresponding simply to the possible value pairs $(b, s) \in \{\underline{b}, \overline{b}\} \times \{\underline{s}, \overline{s}\}$. For each state, the table shows the

$\overline{s}$	$p_{\underline{b}}p_{\overline{s}}$ (η_B^1, η_S^1)	$p_{\overline{b}}p_{\overline{s}}$ (η_B^3, η_S^2)
$\underline{s}$	$p_{\underline{b}}p_{\underline{s}}$ (η_B^2, η_S^3)	$p_{\overline{b}}p_{\underline{s}}$ (η_B^3, η_S^4)
	$\underline{b}$	$\overline{b}$

$\overline{s}$	$p_{\underline{b}}p_{\overline{s}}$ (η_B^4, η_S^5)	$p_{\overline{b}}p_{\overline{s}}$ (η_B^6, η_S^6)
$\underline{s}$	$p_{\underline{b}}p_{\underline{s}}$ (η_B^5, η_S^7)	$p_{\overline{b}}p_{\underline{s}}$ (η_B^7, η_S^7)
	$\underline{b}$	$\overline{b}$

Table 3: Information structures for parameter regions I and II: $\mathcal{S}^1$ (left) and $\mathcal{S}^2$ (right).

probability of that state and the pair of signals (η_B, η_S) received by the buyer and seller.

In information structure $\mathcal{S}^1$, the seller knows exactly which state has occurred; the buyer, if her value is $\overline{b}$, does not know whether the seller is $\underline{s}$ or $\overline{s}$. Information structure $\mathcal{S}^2$ is similar but now it is the low-value seller $\underline{s}$ who is imperfectly informed.

- Second, actually $\mathcal{S}^1$ will serve as our bad information structure for types $\overline{b}$ and $\overline{s}$; $\mathcal{S}^2$ will serve as the bad information structure for types $\underline{b}$ and $\underline{s}$. This is not actually too different from our original hypothesis that $\mathcal{S}^1$ is bad for the buyer and $\mathcal{S}^2$ bad for the seller: intuitively, $\overline{b}$ and $\underline{s}$ are the types whose information matters because they are the ones that earn information rents; we would expect an optimal mechanism to give types $\underline{b}$ and $\overline{s}$ zero payoffs regardless of the information structure.

We summarize the analysis of this parameter region thus:

Lemma 5.1. *If $p_{\overline{b}} \leq p^*_{\overline{b}}$ and $p_{\underline{s}} \leq p^*_{\underline{s}}$, then for the list of information structures shown in Table 3, any list of direct mechanisms $\mathcal{L}$ satisfies*

$$TMU(\mathcal{L}) \leq p_{\overline{b}}(\overline{b} - \overline{s}) + p_{\underline{s}}(\underline{b} - \underline{s}).$$

The proof, in the appendix, indeed shows that the upper bound holds even if we "pre-assign" each type to a particular information structure: using $\mathcal{S}^1$ to evaluate the utility of types $\overline{b}, \overline{s}$, and $\mathcal{S}^2$ for $\underline{b}, \underline{s}$ (rather than having to evaluate each type's payoff by the worst information structure, which may depend on the choice of mechanisms).

5.2 Region II (and II′)

In parameter region II, $p_{\overline{b}} > p^*_{\overline{b}}$, but still $p_{\underline{s}} \leq p^*_{\underline{s}}$.

This region can be analyzed using the same information structures in Table 3. However, we cannot pre-assign each type to one information structure: In particular, for types $\overline{b}$ and $\overline{s}$, we use both information structures in order to bound their minimum utilities from any mechanism list $\mathcal{L}$. (Appendix E.1 shows that pre-assigning types to information structures as before is not sufficient to prove the bound.)

The result of our analysis of this region is as follows:

Lemma 5.2. *If $p_{\overline{b}} > p^*_{\overline{b}}$ and $p_{\underline{s}} \leq p^*_{\underline{s}}$, then for the list of information structures shown in Table 3, any list of direct mechanisms $\mathcal{L}$ satisfies*

$$TMU(\mathcal{L}) \leq p_{\overline{b}}p_{\overline{s}}(\overline{b} - \overline{s}) + p_{\overline{b}}p_{\underline{s}}(\overline{b} - \underline{s}) + p_{\underline{b}}p_{\underline{s}}(\underline{b} - \underline{s})\frac{\overline{b} - \overline{s}}{\overline{b} - \underline{b}}.$$

Note that the bound given in the lemma is indeed the welfare from the optimal dominant-strategy mechanism (Lemma 2.1; note case (b) applies).

We get a corresponding result for region II′. We do not write out a proof since it is completely symmetric.

Lemma 5.3. *If $p_{\underline{s}} > p^*_{\underline{s}}$ and $p_{\overline{b}} \leq p^*_{\overline{b}}$, then for the list of information structures shown in Table 3, any list of direct mechanisms $\mathcal{L}$ satisfies*

$$TMU(\mathcal{L}) \leq p_{\underline{b}}p_{\underline{s}}(\underline{b} - \underline{s}) + p_{\overline{b}}p_{\underline{s}}(\overline{b} - \underline{s}) + p_{\overline{b}}p_{\overline{s}}(\overline{b} - \overline{s})\frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}}.$$

5.3 Region III (and III′)

In parameter region III, $p_{\underline{s}} > p^*_{\underline{s}}$, and $p_{\underline{b}}p_{\underline{s}}\frac{\underline{b}-\underline{s}}{\overline{b}-\underline{b}} \leq p_{\overline{b}}p_{\overline{s}}\frac{\overline{b}-\overline{s}}{\overline{s}-\underline{s}}$ (case (b) of Lemma 2.1).

For this region, the information structures used previously will no longer suffice (Appendix E.2 gives a counterexample). We use the same $\mathcal{S}^1$ as before, but a new $\mathcal{S}^2$, with the following form: Conditional on the buyer having high value $\overline{b}$, with a certain probability $1 - \lambda$, the seller is fully informed and the buyer uninformed (and both parties know this). The rest of the state space is as in the earlier $\mathcal{S}^2$: the buyer is fully informed, and the seller is uninformed if her value is $\underline{s}$.

Here the value of λ is given by

$$\lambda = \frac{p_{\underline{b}}}{p_{\overline{b}}} \times \frac{\underline{b} - \underline{s}}{\overline{b} - \underline{b}}.$$

$\overline{s}$	$p_{\underline{b}}p_{\overline{s}}$ (η_B^1, η_S^1)	$p_{\overline{b}}p_{\overline{s}}$ (η_B^3, η_S^2)
$\underline{s}$	$p_{\underline{b}}p_{\underline{s}}$ (η_B^2, η_S^3)	$p_{\overline{b}}p_{\underline{s}}$ (η_B^3, η_S^4)
	$\underline{b}$	$\overline{b}$

$\overline{s}$	$p_{\underline{b}}p_{\overline{s}}$ (η_B^4, η_S^5)	$\lambda p_{\overline{b}}p_{\overline{s}}$ (η_B^6, η_S^6)	$(1-\lambda)p_{\overline{b}}p_{\overline{s}}$ (η_B^3, η_S^2)
$\underline{s}$	$p_{\underline{b}}p_{\underline{s}}$ (η_B^5, η_S^7)	$\lambda p_{\overline{b}}p_{\underline{s}}$ (η_B^7, η_S^7)	$(1-\lambda)p_{\overline{b}}p_{\underline{s}}$ (η_B^3, η_S^4)
	$\underline{b}$	$\overline{b}$	$\overline{b}'$

Table 4: Information structures for parameter region III: $\mathcal{S}^1$ (left) and $\mathcal{S}^2$ (right).

Note that the assumptions of this parameter region ensure that

$$0 < \lambda \leq \frac{p_{\overline{s}}}{p_{\underline{s}}} \times \frac{\overline{b} - \overline{s}}{\overline{s} - \underline{s}} = \frac{1 - p_{\underline{s}}}{p_{\underline{s}}} \times \frac{p^*_{\underline{s}}}{1 - p^*_{\underline{s}}} < 1.$$

Explicitly, the new information structures are as shown in Table 4. $\mathcal{S}^2$ now consists of six states, labeled by pairs in $\{\underline{b}, \overline{b}, \overline{b}'\} \times \{\underline{s}, \overline{s}\}$; the buyer has high value $\overline{b}$ at states labeled with either $\overline{b}$ or $\overline{b}'$. The table again shows the probability of each state, and the signals received by the buyer and seller. The $\overline{b}'$ states are the ones where the buyer does not know the seller's value.

In particular, the two information structures form an overlapping list: the new $\overline{b}'$ states in $\mathcal{S}^2$ correspond to the $\overline{b}$ states in $\mathcal{S}^1$. Specifically, signals $\eta_B^3, \eta_S^2, \eta_S^4$ are shared between the two information structures.

We thus employ the overlapping version of the extortion lemma. We can pre-assign types $\overline{b}, \overline{s}$ to $\mathcal{S}^1$ and $\underline{b}, \underline{s}$ to $\mathcal{S}^2$, but use the fact that the seller's payoff under signal profile (η_B^3, η_S^4) is the same in $\mathcal{S}^1$ as in $\mathcal{S}^2$. The result is then a bound equal to the welfare from the optimal dominant-strategy mechanism (case (b) of Lemma 2.1):

Lemma 5.4. *If $p_{\underline{s}} > p^*_{\underline{s}}$ and $p_{\underline{b}}p_{\underline{s}}\frac{\overline{b}-\underline{s}}{\overline{b}-\underline{b}} \leq p_{\overline{b}}p_{\overline{s}}\frac{\overline{b}-\overline{s}}{\overline{s}-\underline{s}}$, then for the overlapping list of information structures shown in Table 4, any overlapping list of direct mechanisms $\mathcal{L}$ satisfies*

$$TMU(\mathcal{L}) \leq p_{\overline{b}}p_{\overline{s}}(\overline{b} - \overline{s}) + p_{\overline{b}}p_{\underline{s}}(\overline{b} - \underline{s}) + p_{\underline{b}}p_{\underline{s}}(\underline{b} - \underline{s})\frac{\overline{b} - \overline{s}}{\overline{b} - \underline{b}}.$$

Again, there is a corresponding result for region III′, which we give without a separate proof. For brevity we avoid explicitly writing out the overlapping list of information structures (it is the mirror image of the one in Table 4).

Lemma 5.5. *If $p_{\overline{b}} > p^*_{\overline{b}}$ and $p_{\underline{b}}p_{\underline{s}}\frac{\overline{b}-\underline{s}}{\overline{b}-\underline{b}} \geq p_{\overline{b}}p_{\overline{s}}\frac{\overline{b}-\overline{s}}{\overline{s}-\underline{s}}$, then there is an overlapping list of*

information structures such that any corresponding list $\mathcal{L}$ of overlapping direct mechanisms satisfies

$$TMU(\mathcal{L}) \leq p_{\underline{b}} p_{\underline{s}}(\underline{b} - \underline{s}) + p_{\overline{b}} p_{\underline{s}}(\overline{b} - \underline{s}) + p_{\overline{b}} p_{\overline{s}}(\overline{b} - \overline{s}) \frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}}.$$

As a side note, the overlapping version of the extortion lemma is needed: Appendix E.3 gives a counterexample to show that the information structures used here give only a weaker bound if the overlapping requirement is dropped.

6 Discussion of techniques

In this section we address a number of loose ends. In particular, we consider several technical complexities that arise in the proofs, and discuss whether they can be simplified.

6.1 Free information

Our main result showed that, if a designer desires robustness to *costly* information games, she can do no better than either a dominant-strategy or a flexible-price mechanism.

Can we do away with costly information acquisition? In particular, what if we restrict to information games in which signals simply arrive exogenously at no cost. Thus the designer seeks robustness to unknown information structures, as in the optimal auction problem of Brooks and Du [7]. Does this already hold the designer down to our same welfare guarantee? If so, then we would have no need for all the apparatus of extortion games, and the characterizations in Section 5 could also be made simpler since we would be able to use a single worst-case information structure instead of a list.

We show here that this is not the case: At least for some parameters, if the designer only seeks robustness to information that arrives for free, she can guarantee welfare strictly higher than that in Theorems 3.2 and 3.3. For brevity we keep the exposition slightly less formal.

Specifically, we will focus on parameters with $p_{\overline{b}} < p^*_{\overline{b}}$ and $p_{\underline{s}} < p^*_{\underline{s}}$, so that the original theorem identifies a flexible-price mechanism as optimal, and W_{FP} as the corresponding welfare guarantee. Here is an intuition for how to improve when information is free: Consider the flexible-price mechanism, with the seller offering. The worst information structure for this mechanism is one where, whenever the type realization is $(\overline{b}, \underline{s})$, the low-type seller receives a noisy signal that makes her close to indifferent as to which price to offer, so that she just barely prefers to offer the high price $\overline{s}$. Indeed, for any other

information structure, either the low-type seller strictly benefits from her information and so earns payoff above $\underline{b} - \underline{s}$, or else the high-type buyer sometimes gets offered the low price $\underline{b}$ and so earns payoff above $\overline{b} - \overline{s}$.

Then, we can achieve an improved welfare by randomizing between the flexible-price mechanism and some other mechanism that does better on this particular information structure.

Specifically: Fix any positive number ϵ such that

$$1 - \frac{p_{\underline{s}}}{p_{\overline{s}}} \times \frac{\overline{s} - \underline{b}}{\overline{b} - \overline{s}} - \frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}} < \epsilon < 1 - \frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}}.$$

(The expression on the left is positive; this follows from $p_{\underline{s}} < p^*_{\underline{s}}$.) Consider the following mechanism $\mathcal{M}_\delta$, where $\delta \in (0, 1)$ is a parameter:

(i) With probability $1-\delta$, run the flexible-price mechanism: The seller can choose price $\underline{b}$ or $\overline{s}$. The buyer can then accept or reject.

(ii) With remaining probability δ, run the following mechanism: The seller can choose either to offer trade at price $\underline{b}$, or a lottery in which, with probability $q_\epsilon = (\underline{b} - \underline{s})/(\overline{s} - \underline{s}) + \epsilon$, trade occurs at price $\overline{s}$. The buyer can then accept or reject. (This is the mechanism described in the proof of Proposition 2.2, in Appendix D).

For any information structure $\mathcal{S}$, consider the information game $\mathcal{I}(\mathcal{S})$ where the players take no actions, and simply receive signals according to $\mathcal{S}$ at no cost. We then have:

Proposition 6.1. *For any sufficiently small δ, the infimum of $\overline{W}(\mathcal{M}_\delta, \mathcal{I}(\mathcal{S}))$ over all information structures $\mathcal{S}$ is strictly higher than W_{FP}.*

The proof is in Appendix D.

Note that we could again obtain a corresponding result using the welfare criterion $\underline{W}$, by slightly elaborating the mechanism to get rid of spurious undominated strategies.

6.2 The role of informants

An awkward feature of the construction of the worst-case information game, under the best-equilibrium criterion $\overline{W}$, is the introduction of additional players I_B, I_S to provide information about what equilibrium is being played. Could we get rid of these additional players? In particular, could we have the original players B, S themselves supply the needed information?

It would be natural to consider a construction in which B reports the seller's least-preferred information structure and willingness-to-pay, $\underline{k}(s)$ and Δ_s, and simultaneously S reports $\underline{k}(b)$ and Δ_b. However, a difficulty that any such construction must confront is the possibility of equilibria where the low type of buyer $\underline{b}$ reports a different choice of $(\underline{k}(s), \Delta_s)$ than the high type $\overline{b}$ does. Then, when the seller observes the extortion offer, she can make inferences about the buyer's value, and so continuation play after some information structure $\mathcal{S}^k$ is realized does not need to correspond to a direct mechanism on $\mathcal{S}^k$ (because each player may already have additional information about the other). This informational difference also means that the seller's willingness-to-pay after receiving one extortion offer is different than her willingness-to-pay after a different offer, so that it may indeed be possible to simultaneously sustain two different offers in equilibrium, with each offer being a truthful report of the seller's willingness-to-pay after receiving that offer.

A related challenge is the possibility of mixed equilibria, where — even conditional on a given value, say $\overline{b}$ — the buyer may mix between different reports $(\underline{k}(s), \Delta_s)$, and then correlate her reports with her actions in the mechanism. This correlation again can make it possible to simultaneously sustain multiple different extortion offers.

Supposing that we cannot get rid of extra players, can we at least use a single informant, instead of having two separate informants I_B and I_S? That is, can we have one player I who reports both the buyer's and the seller's willingnesses-to-pay? Here again, it seems difficult to rule out equilibria that involve mixing and correlation in the reports.

In brief, the key property of our construction, with separate I_B and I_S, is that the buyer does not incidentally learn anything during the information game that could possibly be informative about the seller's subsequent play in the mechanism (nor vice versa). This ensures that when signals (η_B^k, η_S^k) are realized, the players indeed have learned nothing about each other beyond what is specified in the information structure $\mathcal{S}^k$. It is not clear how to do away with informants without sacrificing this key property.

6.3 On choice of information structures

The proof of our main result involved several different parameter regions and choices of information structures to apply the extortion lemma. Is this complexity necessary?

In Appendix E, we consider this question more extensively, by examining several ways one might try to simplify the arguments, and giving counterexamples to show that these attempts fail. Here, for brevity, we simply summarize the findings (which have also been

mentioned earlier):

- In parameter region I (subsection 5.1), we were able to "pre-assign" each type to a particular information structure, and use only its payoff in that information structure to obtain the overall bound on $TMU(\mathcal{L})$. In parameter region II, even though the same list of information structures is used, we cannot pre-assign types in the same way. Appendix E.1 gives a counterexample.

- In parameter region III, we used a different, and more complex, pair of information structures than in parameter regions I and II. Appendix E.2 gives an example to show that using the information structures from regions I and II would not have sufficed to get a tight bound in region III.

- Also in parameter region III, we used the strengthened version of the extortion lemma that considered overlapping lists of direct mechanisms. Appendix E.3 shows by example that the non-overlapping version of the lemma, applied to the same information structures, would not suffice.

6.4 General tightness of the extortion lemma

Our key tool for the analysis was the extortion lemma, which provides an upper bound on the best possible welfare guarantee of any mechanism. As discussed earlier, this lemma can be formulated much more generally than our two-type bilateral trade model. In our analysis, the lemma always gives a tight bound for the optimal mechanism — as we showed by exhibiting a mechanism that achieves the bound.

Is the bound always tight? That is, is it true beyond our setting that there is some appropriate choice of (overlapping) information structures $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ such that $\sup_{\mathcal{M}} W(\mathcal{M}) = \max_{\mathcal{L}} TMU(\mathcal{L})$?

It seems that this should be true: an intuition is that each type of each agent cannot be forced down to a lower payoff than the worst she gets in equilibrium across all possible information structures. However, this is not obviously correct. Conceivably, it might be possible to drive an agent's payoff even lower than the worst information structure for her, by a variation of the extortion construction we have used. Notice that the amount that can extorted from the buyer, in our model, is the difference between her payoff on the equilibrium path (where she accepts extortion) and off the path (where she rejects). We have used information structure $\mathcal{S}^{\underline{k}(b)}$ to generate the off-path payoffs for the buyer (and likewise for the seller). But in principle we could instead use an off-path continuation that

"deceives" the seller, by sending her signals η_S^k that come from some information structure $\mathcal{S}^k$ but do not follow the distribution in π^k. Such a continuation could potentially be even worse for the buyer than any actual information structure $\mathcal{S}^k$, and thereby allow us to construct an adversarial information game that extorts from the buyer an amount even greater than Δ_b. So, sharpening the extortion lemma to a tight bound in general would require showing that no such construction is possible.

7 Conclusion

We have considered a broad class of information games, in which agents can take costly actions to acquire information or influence others' information (or both), and asked how a planner might choose among trading mechanisms, with an eye to their implications for the costs spent on such activities. We formalized the planner's overall objective as expected welfare, which includes both the surplus generated within the trading mechanism itself and the costs incurred in the information game. Rather than make some (necessarily arbitrary) assumption about the particular information game available, we have considered all such games, and used a worst-case robustness criterion to evaluate welfare. We focused on a bilateral trade model as a natural case study. We have considered only the simplest case — two types of each agent — but this allows for a complete and clean characterization.

In this setting (or any other private-values setting), dominant-strategy mechanisms are a natural class of candidates for the optimally robust mechanism, since they give no incentive for agents to manipulate information in any way. However, in our problem, the optimum is not always a dominant-strategy mechanism. Moreover, when a dominant-strategy mechanism is not optimal, the optimum is instead a quite simple mechanism, where one agent can choose a price to offer to the other. And the parameter ranges for preferring one mechanism or the other are also intuitive: Consider the state where both agents' values are low and the state where both values are high; no dominant-strategy mechanism can realize all gains from trade in both these cases. Thus, if both of these situations are likely (relative to the other case where trade is desirable — high buyer value and low seller value), then it is better to use a more flexible mechanism. In this case, each agent's possible incentives to influence information are small, so any costs of such influence are outweighed by the allocative benefits of the flexible-price mechanism.

The analysis also led to a simple description of the worst-case information game. It is not one where each player pays to acquire information, but rather one where each player must pay to *prevent* information from being released in an undesirable way. The particular

information structure that is bad for each player (and each type) is not predetermined but is endogenous to the mechanism, and to the choice of equilibrium being played. This idea led to the extortion lemma, a tool for upper-bounding the informationally robust welfare in any mechanism, which can be applied beyond the setting we have studied: it essentially says that no mechanism can guarantee any better than giving each type its worst payoff across all information structures.

What are next steps to take? There are two natural directions for improvement. First, it would be desirable to find a way to avoid introducing extra players in the information game. This would make for a simpler theorem statement and a more technically satisfying analysis. And, second, it would of course be desirable to extend the results to give a tight analysis for arbitrarily many types of buyer and seller, and to give techniques for other mechanism design problems as well.

The extortion lemma gives an upper bound for welfare in any setting, and it is natural to expect that this same technical machinery can be used in other mechanism design problems to likewise identify the optimally robust mechanism. The trick is that getting a tight bound requires a judicious choice of information structures to apply the lemma. In the analysis here, this choice was made by hand. It appears that finding the right way to make this choice in general will be a key step in order to extend the analysis to other problems. Even when finding the optimal mechanism is difficult, however, the simple insights about the flexible-price mechanism obtained here may prove useful in finding some robust improvements over dominant-strategy mechanisms in other applications.

A Full definition of information games

Here we lay out the formal definition of information games. For the reasons introduced in Section 3, we will allow for additional players besides B and S. The definition is based on the standard formulation of extensive-form games, as in [13]. We allow for moves of nature. Our definition would be slightly more compact if we required nature to move only once at the beginning (as in [16]), but we will find it convenient to allow multiple moves of nature; the two formulations are substantively equivalent.

An *information game* consists of a pair $\mathcal{I} = (\mathcal{P}, \mathcal{G})$, where $\mathcal{P} = (\Omega, \pi, b, s)$ is a probability space, and $\mathcal{G}$ consists of the following objects:

- A set of players $N = \{1, 2, \ldots, n\}$, where $n \geq 2$ (we will take player 1 to be the buyer and 2 to be the seller, and refer to them also as B and S). We will use the

term *information game without additional players* to refer to the case $n = 2$ (and *with additional players allowed* to emphasize the general case).

- A finite rooted tree. We write X for the set of nonterminal nodes and Z for the set of terminal nodes, and x_0 for the initial node.
- A move function $\iota : X \to N \cup \{0\}$, specifying who moves at each node (or 0 for nature).
- Information partitions for each player, $(H_1, \ldots, H_n)$: specifically, for $i = 1, 2$, H_i (with typical element h_i) is a partition of $\iota^{-1}(i) \cup Z$, and for each $i > 2$, H_i is a partition of $\iota^{-1}(i)$.
- For each i and each $h_i \in H_i$, a set of action labels $A(h_i)$.
- For each node $x \in X$ with $\iota(x) = i \neq 0$, a bijection from the action labels $A(h_i(x))$ to the successors of x in the tree.

  (Note, this ensures that a terminal node and a nonterminal node can never be in the same information set, since a terminal node must lie in an information set whose set of action labels is empty.)
- For each node x with $\iota(x) = 0$, a function from Ω to the successors of x.
- For each $i \in N$, a payoff function $g_i : Z \to \mathbb{R}$.

We further require these objects to satisfy conditions (a)–(d) below. To state these conditions, we first restate some standard definitions in our context.

A *pure strategy* s_i for player $i \in N$ is a function that picks out, for each information set $h_i \in H_i$ consisting of nonterminal nodes, an action $s_i(h_i) \in A(h_i)$. Given a pure strategy for each player and a state $\omega \in \Omega$, these together pick out a unique path through the tree (and in particular a unique terminal node). We say a node x is *reachable under strategy* s_i if there exist strategies for the other players and a state ω such that x is on the path of play. We also say x is *reachable in state* ω if there exist strategies for each player such that x is on the path of play under state ω.

At a node x with $\iota(x) = i$, the *experience* of player i is the sequence of previous information sets of player i and actions played there, on the unique path from the start node to x [17].

Our requirements are as follows:

(a) (Perfect recall) For any player i, any $h_i \in H_i$ and two nodes $x, x' \in h_i$, the experience of i at x is the same as at x'.

(b) (No redundant nodes) For every node, there is some state in which it is reachable. (Note that this is a requirement on the moves of nature.)

(c) (Known own values) If x, x' are two nodes in the same information set of player B, reachable in states ω and ω' respectively, then $b(\omega) = b(\omega')$. If x, x' are two nodes in the same information set of player S, reachable in states ω and ω' respectively, then $s(\omega) = s(\omega')$.

(d) (Inaction strategies) For every player $i \in N$, there exists a pure strategy s_i such that $g_i(z) \geq 0$ at every terminal node z reachable under s_i.

This is also a natural place to describe explicitly how an information game and a mechanism come together to form a combined game. Suppose an information game $\mathcal{I} = (\mathcal{P}, \mathcal{G})$ as above is given, and $\mathcal{M} = (A_B, A_S, q, t)$ is a mechanism. We would like to specify that the players play $\mathcal{I}$ and then $\mathcal{M}$. But $\mathcal{M}$ is described with simultaneous moves, and the usual formulation of extensive-form games requires one player to move at a time; arbitrarily we will specify that B moves before S. So the combined game is an extensive-form game derived from $\mathcal{G}$ as follows:

- The players are the same as in $\mathcal{G}$.
- The nodes consist of the nodes of $\mathcal{G}$, plus additional nodes of the form (z, a_B) and (z, a_B, a_S) for each $z \in Z$, $a_B \in A_B$, $a_S \in A_S$. We specify that each node (z, a_B, a_S) is a successor of (z, a_B), which is in turn a successor of z. (Thus the terminal nodes of the combined game are the nodes of the form (z, a_B, a_S).)
- Player B moves at each node $z \in Z$, and S moves at each node of the form (z, a_B).
- Player B's information sets are the same as in $\mathcal{G}$. Player S has the same information sets as in $\mathcal{G}$, plus one information set of the form $h_S \times A_B$ for each information set h_S consisting of terminal nodes in $\mathcal{G}$. Each other player's information sets are the same as in $\mathcal{G}$.
- Actions are the same as in $\mathcal{G}$, where applicable. At each information set for player B consisting of terminal nodes in $\mathcal{G}$, the set of actions is A_B; at each information set for player S of the form $h_S \times A_B$, the set of actions is A_S. The mapping from actions at these information sets to successor nodes is the obvious one.

- Moves of nature are the same as in $\mathcal{G}$.
- Payoff functions for B and S over terminal nodes are defined by

$$\begin{aligned} u_B(z, a_B, a_S) &= g_B(z) + b(z)q(a_B, a_S) - t(a_B, a_S), \\ u_S(z, a_B, a_S) &= g_S(z) + t(a_B, a_S) - s(z)q(a_B, a_S), \end{aligned}$$

  where $b(z)$ is defined to be $b(\omega)$ for any ω under which z is reachable (note this is well-defined by conditions (b) and (c)), and similarly for $s(z)$.

  Payoff functions for any other player $i > 2$ are defined by

$$u_i(z, a_B, a_S) = g_i(z).$$

This is a standard extensive-form game, aside from the fact that moves of nature at any node x are still referenced by states. We can easily convert each such description to a probability distribution over x's successors, by using the distribution π conditional on the set of states under which x is reachable. Note that assumption (b) ensures that each such distribution is well-defined and has full support.

This fully defines the combined game, in the usual framework of extensive-form games with perfect recall, and we can apply notions such as sequential equilibrium.

B Proofs for extortion lemma

For convenience, this appendix is divided into four sections — first a description of the adversarial information game, and then the actual proof of the welfare bound, for each of the two versions of the extortion lemma (worst-equilibrium and best-equilibrium). Throughout, we take the mechanism $\mathcal{M} = (A_B, A_S, q, t)$ as fixed, as well as the overlapping list of information structures $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$, and an arbitrary default information structure $\mathcal{S}^0$. We assume the signal sets in $\mathcal{S}^0$ to be disjoint from those of each other $\mathcal{S}^k$, and thereby can view $(\mathcal{S}^0, \ldots, \mathcal{S}^K)$ as an overlapping list.

We will describe the information games without explicit reference to the probability spaces, but it should be apparent that the games can be constructed as described, using an appropriate common refinement of the Ω^k for the underlying probability space.

B.1 The information game (worst-equilibrium criterion)

First consider the following auxiliary game: An index $k \in \{0, \ldots, K\}$ is drawn uniformly at random; the buyer and seller then receive signals (η_B, η_S) drawn according to information structure $\mathcal{S}^k$ (without being informed of the choice of k); then they play the mechanism $\mathcal{M}$. Consider any undominated-strategy equilibrium of this game.

This leads to an overlapping list of direct mechanisms via the revelation principle. Specifically, for each possible signal η_B, let $\alpha_B(\eta_B)$ denote the buyer's action in the mechanism after receiving signal η_B, and similarly $\alpha_S(\eta_S)$ for the seller. (Note that we need not further condition on the players' values, since these are uniquely determined by their signals.) Then, for each k and $(\eta_B, \eta_S) \in H_B^k \times H_S^k$, define $q^k(\eta_B, \eta_S) = q(\alpha_B(\eta_B), \alpha_S(\eta_S))$ and $t^k(\eta_B, \eta_S) = t(\alpha_B(\eta_B), \alpha_S(\eta_S))$, where q and t are extended to mixed actions by linearity.

Lemma B.1. *The mechanisms $\mathcal{M}^k = (q^k, t^k)$ form an overlapping list of direct mechanisms, for the given information structures.*

The proof is is Appendix D. This list will be our $\mathcal{L}$ such that $\underline{W}(\mathcal{M}) \leq TMU(\mathcal{L})$.

Now define

$$\underline{k}(\underline{b}) = \arg\min_{k \geq 1} u_{\underline{b}}(\mathcal{M}^k), \qquad \Delta_{\underline{b}} = u_{\underline{b}}(\mathcal{M}^0) - u_{\underline{b}}(\mathcal{M}^{\underline{k}(\underline{b})}),$$

$$\underline{k}(\overline{b}) = \arg\min_{k \geq 1} u_{\overline{b}}(\mathcal{M}^k), \qquad \Delta_{\overline{b}} = u_{\overline{b}}(\mathcal{M}^0) - u_{\overline{b}}(\mathcal{M}^{\underline{k}(\overline{b})}),$$

$$\underline{k}(\underline{s}) = \arg\min_{k \geq 1} u_{\underline{s}}(\mathcal{M}^k), \qquad \Delta_{\underline{s}} = u_{\underline{s}}(\mathcal{M}^0) - u_{\underline{s}}(\mathcal{M}^{\underline{k}(\underline{s})}),$$

$$\underline{k}(\overline{s}) = \arg\min_{k \geq 1} u_{\overline{s}}(\mathcal{M}^k), \qquad \Delta_{\overline{s}} = u_{\overline{s}}(\mathcal{M}^0) - u_{\overline{s}}(\mathcal{M}^{\underline{k}(\overline{s})}).$$

(In each case, if there is more than one minimizing k then choose one arbitrarily.)

We can now describe our extortion information game. Let $\epsilon > 0$ be small. Initially, players know only their own values. Then:

1. The buyer chooses A or R (accept or reject extortion). Simultaneously, the seller chooses A or R.

2. A value of k, together with payoffs g_B, g_S, are determined as follows.

   - With exogenous probability ϵ, the value of $k \in \{0, 1, \ldots, K\}$ is chosen uniformly at random, and g_B and g_S are both zero.

– Otherwise: If the buyer chose R, then k is $\underline{k}(\underline{b})$ or $\underline{k}(\overline{b})$ (depending on the buyer's value in the realized state), and $g_B = g_S = 0$.

If the buyer chose A but the seller chose R, then k is $\underline{k}(\underline{s})$ or $\underline{k}(\overline{s})$ (depending on the seller's value), and $g_B = -\Delta_{\underline{b}}$ or $-\Delta_{\overline{b}}$ (depending on the buyer's value) while $g_S = 0$.

If both parties chose A, then $k = 0$, $g_B = -\Delta_{\underline{b}}$ or $-\Delta_{\overline{b}}$, and $g_S = -\Delta_{\underline{s}}$ or $-\Delta_{\overline{s}}$.

3. Signals (η_B, η_S) are drawn according to the corresponding information structure $\mathcal{S}^k$. The parties observe their signals (but do not further observe the realized k). Payoffs are g_B and g_S as above.

This game meets the inaction requirement: each player can guarantee herself 0 by choosing R.

It will be useful to describe the terminal nodes and the players' knowledge about them. The possible terminal nodes are of the form $(d_B, d_S, d_0, k, \eta_B, \eta_S)$ (where $d_B, d_S \in \{A, R\}$ indicate the buyer's and seller's responses to extortion, and d_0 indicates which of the two cases in step 3 occurred). The buyer's information consists of (d_B, η_B). (We need not further include the value b in the buyer's information since it is determined by η_B.) The seller's information consists of (d_S, η_S).

B.2 Proof of lemma (worst-equilibrium criterion)

Proof of Lemma 4.3. We have constructed a list of direct mechanisms $\mathcal{L}$, in the preceding subsection. We will show that, with the specified information game $\mathcal{I}$, the combined game has an undominated sequential equilibrium whose welfare is close to $TMU(\mathcal{L})$. Let $\alpha_B(\eta_B)$ and $\alpha_S(\eta_S)$ denote the equilibrium actions from the auxiliary game.

We first describe the strategies in the combined game: The buyer, when asked to choose A or R, always chooses A; the seller also always chooses A. When asked to play the mechanism, if the buyer has received some signal η_B, she acts according to $\alpha_B(\eta_B)$; similarly, the seller acts according to $\alpha_S(\eta_S)$.

Under these strategies, any information set of either player can be reached without a deviation by the other player (recall the ϵ chance of random signal structure in stage 2). Hence, beliefs are uniquely determined by Bayesian updating. Thus consistency of the resulting assessment is automatic, and we just need to check sequential rationality.

Consider any information set of the buyer at the mechanism stage, either on or off path (so d_B may have been either A or R). What is the buyer's belief about η_S? Note that

because both the low- and high-value seller are expected to play $d_S = A$ for certain, the distribution over terminal nodes of the information game, given d_B, can be decomposed in the form

$$\rho(d_0, k | d_B, A) \times \pi^k(\eta_B, \eta_S)$$

(where ρ describes how (d_0, k) are determined at stage 2 given (d_B, d_S)). In particular, conditional on k and on η_B, the belief about η_S is distributed $\pi^k(\eta_S|\eta_B)$, which we know can be written as $\pi(\eta_S|\eta_B)$ without reference to k. Hence, this is also the buyer's belief about the seller's signal, unconditional on k. (Note that this argument depends on the fact that both the low- and high-value seller play A: otherwise, the realized η_B could be informative about k, therefore about d_S, and thereby give additional information about the seller's type.) Consequently, the buyer expects the seller to act in the mechanism according to $\sum_{\eta_S} \pi(\eta_S|\eta_B)\alpha_S(\eta_S)$. Then α_B is indeed a best reply; this follows from equilibrium in the auxiliary game.

Similarly, since the buyer is always expected to play $d_B = A$, the seller's beliefs after playing d_S can be expressed in the form

$$\rho(d_0, k | A, d_S) \times \pi^k(\eta_B, \eta_S).$$

The belief about η_B conditional on k and η_S is given by $\pi^k(\eta_B|\eta_S) = \pi(\eta_B|\eta_S)$, and so this is the belief about η_B given η_S alone. So the seller expects the buyer to play $\sum_{\eta_B} \pi(\eta_B|\eta_S)\alpha_B(\eta_B)$, and so $\alpha_S(\eta_S)$ is indeed a best reply.

What about optimality of the stage-1 accept/reject decision? Note that if, in stage 2, a random k is chosen, then the stage-1 decision has no effect on subsequent play or payoffs. Conditioning on the complementary case in stage 2, and on the buyer's value $b \in \{\underline{b}, \overline{b}\}$, the buyer's expected payoff if she chooses A in stage 1 is $-\Delta_b$ in the information game and $u_b(\mathcal{M}^0)$ in the mechanism, for a total payoff of $-\Delta_b + u_b(\mathcal{M}^0) = u_b(\mathcal{M}^{\underline{k}(b)})$. If she instead chooses R, the corresponding expected payoff is 0 in the information game and $u_b(\mathcal{M}^{\underline{k}(b)})$ in the mechanism. (These calculations use the fact that the seller is playing $d_S = A$.) Hence the buyer is indifferent.

Similarly, consider the seller's stage-1 decision. Again, if a random k is chosen in stage 2, the stage-1 decision has no effect. Conditioning on the complementary case in stage 2, and on the seller's value $s \in \{\underline{s}, \overline{s}\}$: if the seller chooses A then the conditional expected payoff is $-\Delta_s$ in the information game and $u_s(\mathcal{M}^0)$ in the mechanism, hence $-\Delta_s + u_s(\mathcal{M}^0) = u_s(\mathcal{M}^{\underline{k}(s)})$ in total; if she chooses R, she gets 0 in the information game and $u_s(\mathcal{M}^{\underline{k}(s)})$ in the mechanism. So the seller is also indifferent.

This verifies sequential rationality of the proposed strategies. And the fact that the actions played in the mechanism are undominated is imediate from the original choice of α_B, α_S. Thus we have an undominated sequential equilibrium.

What is the expected total welfare? Conditional on a random k in stage 2, the total welfare is

$$W(\mathcal{M}^k) = p_{\underline{b}} u_{\underline{b}}(\mathcal{M}^k) + p_{\overline{b}} u_{\overline{b}}(\mathcal{M}^k) + p_{\underline{s}} u_{\underline{s}}(\mathcal{M}^k) + p_{\overline{s}} u_{\overline{s}}(\mathcal{M}^k).$$

Conditional on the nonrandom outcome in stage 2, the above calculation shows that the buyer's (combined) payoff is $u_b(\mathcal{M}^{\underline{k}(b)}) = u_b(\mathcal{L})$, and the seller's is $u_s(\mathcal{L})$; so expected total welfare is $TMU(\mathcal{L})$. Thus, overall, expected total welfare is

$$\epsilon \times \frac{1}{K+1} \sum_{k=0}^{K} W(\mathcal{M}^k) + (1-\epsilon) \times TMU(\mathcal{L}).$$

This is an upper bound on the welfare of the worst equilibrium, $\underline{W}(\mathcal{M}, \mathcal{I})$, and therefore on $\underline{W}(\mathcal{M})$. Now taking $\epsilon \to 0$ gives $\underline{W}(\mathcal{M}) \leq TMU(\mathcal{L})$. The lemma follows. □

B.3 The information game (best-equilibrium criterion)

First, given the mechanism $\mathcal{M}$, let M be a large number: specifically, let $t_{\max}$ and $t_{\min}$ be the maximum and minimum values of t, and take $M > (\overline{b} - \underline{s}) + (t_{\max} - t_{\min})$. Thus M is larger than any possible payoff difference between two outcomes of the mechanism. Let F ("fine") be another arbitrarily large number. Also, let $\epsilon, \delta > 0$ be small numbers. Assume δ is chosen such that M/δ is an integer. Assume ϵ is chosen small enough so that

$$\frac{1-\epsilon}{2} \times (1 - 2\epsilon^3) \geq \frac{1}{4} \qquad \text{and} \qquad \epsilon \leq \frac{(1-2\epsilon^3)\pi_{\min}}{2(K+1)},$$

where $\pi_{\min} > 0$ is such that $\pi^k(\eta_B), \pi^k(\eta_S) \geq \pi_{\min}$ for all information structures $k = 0, \ldots, K$ and all $\eta_B \in H_B^k$, $\eta_S \in H_S^k$. And assume F is large enough such that

$$F \geq \frac{3M}{\epsilon^2 \cdot \min\{p_{\underline{b}}, p_{\overline{b}}, p_{\underline{s}}, p_{\overline{s}}\}}.$$

Our information game, with additional players I_B, I_S, is as follows. Initially, B and S know their values b, s. Then:

1. I_B can either choose inaction ($\emptyset$), or can make a report $(\widehat{\underline{k}}(\underline{b}), \widehat{\underline{k}}(\overline{b}), \widehat{\Delta}_{\underline{b}}, \widehat{\Delta}_{\overline{b}})$, where $\widehat{\underline{k}}(\underline{b}), \widehat{\underline{k}}(\overline{b}) \in \{1, \ldots, K\}$, and $\widehat{\Delta}_{\underline{b}}, \widehat{\Delta}_{\overline{b}} \in \{-M, -M+\delta, -M+2\delta, \ldots, M-\delta, M\}$.

Simultaneously, I_S can either choose inaction ($\emptyset$) or can make a report $(\widehat{\underline{k}}(\underline{s}), \widehat{\underline{k}}(\overline{s}), \widehat{\Delta}_{\underline{s}}, \widehat{\Delta}_{\overline{s}})$, where $\widehat{\underline{k}}(\underline{s}), \widehat{\underline{k}}(\overline{s}) \in \{1, \ldots, K\}$, and $\widehat{\Delta}_{\underline{s}}, \widehat{\Delta}_{\overline{s}} \in \{-M, -M+\delta, \ldots, M\}$.

If I_B chooses $\emptyset$, then we put $\widehat{\underline{k}}(\underline{b}) = \widehat{\underline{k}}(\overline{b}) = 1$ and $\widehat{\Delta}_{\underline{b}} = \widehat{\Delta}_{\overline{b}} = 0$. Similarly if I_S chooses $\emptyset$.

2. The buyer is informed of $(\widehat{\underline{k}}(b), \widehat{\Delta}_b)$ (which is $(\widehat{\underline{k}}(\underline{b}), \widehat{\Delta}_{\underline{b}})$ or $(\widehat{\underline{k}}(\overline{b}), \widehat{\Delta}_{\overline{b}})$ depending on her value). The buyer chooses A or R.

   Simultaneously, the seller is informed of $(\widehat{\underline{k}}(s), \widehat{\Delta}_s)$, and chooses A or R.

3. A value of k, and payoffs g_B, g_S, are determined as follows:

   (a) With exogenous probability ϵ, the value of $k \in \{0, \ldots, K\}$ is chosen uniformly at random, and $g_B = g_S = 0$.

   (b) With probability $(1-\epsilon)/2$:

   If the buyer chose R, then $k = \widehat{\underline{k}}(b)$, and $g_B = g_S = 0$.

   If the buyer chose A but the seller chose R, then $k = \widehat{\underline{k}}(s)$, $g_B = -\widehat{\Delta}_b$, and $g_S = 0$.

   If both chose A, then $k = 0$, $g_B = -\widehat{\Delta}_b$ and $g_S = -\widehat{\Delta}_s$.

   (c) With remaining probability $(1-\epsilon)/2$:

   If the seller chose R, then $k = \widehat{\underline{k}}(s)$, and $g_B = g_S = 0$.

   If the seller chose A but the buyer chose R, then $k = \widehat{\underline{k}}(b)$, $g_B = 0$, and $g_S = -\widehat{\Delta}_s$.

   If both chose A, then $k = 0$, $g_B = -\widehat{\Delta}_b$ and $g_S = -\widehat{\Delta}_s$.

4. Signals (η_B, η_S) are drawn according to the information structure $\mathcal{S}^k$. Players B and S observe their signals (but not the realized k). Their payoffs in the information game are g_B, g_S as above.

5. The informants' payoffs are determined as follows: If I_B chose $\emptyset$, his payoff is 0. Otherwise, his payoff is $\epsilon(\widehat{\Delta}_b + 2M)$ if B chose A, and $\epsilon(\widehat{\Delta}_b + 2M) - F$ if B chose R.

   If I_S chose $\emptyset$, his payoff is 0. Otherwise, his payoff is $\epsilon(\widehat{\Delta}_s + 2M)$ if S chose A, and $\epsilon(\widehat{\Delta}_s + 2M) - F$ if S chose R.

This game meets the inaction requirement: B and S can each guarantee a payoff of 0 by choosing R, and the informants can each guarantee a payoff of 0 by choosing $\emptyset$.

Again, in order to talk about strategies in the combined game, it will be useful to briefly note what information the players have each time they make a decision, and at the end of the information game. The informants move only at the beginning, and they have no information. When the buyer chooses A or R, she knows her value b and the pair $(\widehat{\underline{k}}_b, \widehat{\Delta}_b)$. At the end of the information game, she also knows her own previous action (A or R) and the realized signal η_B. Similarly, when the seller first chooses, she knows s and the pair $(\widehat{\underline{k}}_s, \widehat{\Delta}_s)$. At the end of the information game, she also knows her own previous action and the signal η_S.

B.4 Proof of lemma (best-equilibrium criterion)

The proof is a bit lengthy, so we give an outline. We want to show that any sequential equilibrium of the combined game, using the information game in Subsection B.3, has welfare bounded above by $\max_{\mathcal{L}} TMU(\mathcal{L})$ (plus a small fudge factor). We will proceed in six main steps:

- *Step 1.* Neither informant chooses inaction, in equilibrium.
- *Step 2.* In equilibrium, each type of buyer and seller has very low probability of playing R.
- *Step 3.* Consider any two different information sets h_B, h'_B for the buyer (either on or off the equilibrium path), in which she has received the same signal η_B, and now has to choose an action in the mechanism. Her belief about the seller's information at h_B is approximately the same as her belief about the seller's information at h'_B. (As a consequence, her conjecture at h_B about what the seller will do in the mechanism is approximately the same as at h'_B.)
- *Step 4.* The buyer's equilibrium action in the mechanism, conditional on each signal η_B — and likewise the seller's action conditional on each η_S — can be used to construct overlapping "direct mechanisms" $\mathcal{M}^k$ that are approximately IC and IR.
- *Step 5.* In equilibrium, the informants' reports $\widehat{\Delta}_{\underline{b}}, \widehat{\Delta}_{\overline{b}}, \widehat{\Delta}_{\underline{s}}, \widehat{\Delta}_{\overline{s}}$ are close to the agents' true willingnesses-to-pay, as represented by the difference in their expected payoffs between $\mathcal{M}^0$ and $\mathcal{M}^k$.

- *Step 6.* In equilibrium, the buyer's and seller's expected payoffs (in the combined game) are close to their payoffs in the worst $\mathcal{M}^k$.

Before filling in details, we should formally define the "approximate direct mechanisms" in Step 4. Given an information structure $\mathcal{S}$, and a number $\gamma > 0$, we define a *γ-direct mechanism* to be $\mathcal{M} = (q, t)$ with $q : H_B \times H_S \to [0, 1]$, $t : H_B \times H_S \to \mathbb{R}$, satisfying approximate versions of the IC and IR constraints: for all $\eta_B^*, \eta_B' \in H_B$ and $\eta_S^*, \eta_S' \in H_S$,

$$\sum_{\eta_S} \pi(\eta_S|\eta_B^*) \left([b(\eta_B^*)q(\eta_B^*, \eta_S) - t(\eta_B^*, \eta_S)] - [b(\eta_B^*)q(\eta_B', \eta_S) - t(\eta_B', \eta_S)]\right) \geq -\gamma;$$

$$\sum_{\eta_S} \pi(\eta_S|\eta_B^*) \left[b(\eta_B^*)q(\eta_B^*, \eta_S) - t(\eta_B^*, \eta_S)\right] \geq -\gamma;$$

$$\sum_{\eta_B} \pi(\eta_B|\eta_S^*) \left([t(\eta_B, \eta_S^*) - s(\eta_S^*)q(\eta_B, \eta_S^*)] - [t(\eta_B, \eta_S') - s(\eta_S^*)q(\eta_B, \eta_S')]\right) \geq -\gamma;$$

$$\sum_{\eta_B} \pi(\eta_B|\eta_S^*) \left[t(\eta_B, \eta_S^*) - s(\eta_S^*)q(\eta_B, \eta_S^*)\right] \geq -\gamma.$$

If $(\mathcal{S}^1, \ldots, \mathcal{S}^K)$ is an overlapping list of information structures, we define an *overlapping list of γ-direct mechanisms* on it to be $\mathcal{L} = (\mathcal{M}^1, \ldots, \mathcal{M}^K)$, where each $\mathcal{M}^k = (q^k, t^k)$ is a γ-direct mechanism on $\mathcal{S}^k$, and (4.3) is satisfied whenever applicable.

Now we are ready.

Proof of Lemma 4.4. Let $\mathcal{M} = (A_B, A_S, q, t)$ be the given indirect mechanism. Fix M, F, ϵ, δ as in Subsection B.3, and let the information game be as described there. Fix any sequential equilibrium (σ^*, μ^*) of the resulting combined game.

Step 1. Consider the buyer's decision in stage 2 of the information game, at any information set h_B where she is offered $\widehat{\Delta}_b = -M$. Suppose the buyer's strategy σ_B^* calls for choosing R with positive probability. Consider a deviation to play A, and then play in the mechanism as if she had chosen R (holding fixed the received signal η^B).

This deviation has no effect on the buyer's ultimate payoff if stage 3 of the information game leads to realization (a), or realization (c) and the seller chose R. In both cases, the deviation does not change payoffs in the mechanism (because it does not affect the choice of k, so it is not detected by the seller and so does not affect her payoff; nor does it affect the buyer's information), nor does it change payoffs in the information game. On the other hand, if stage 3 leads to realization (b), or realization (c) and the seller chose A, then the deviation changes the buyer's payoff in the information game from 0 to $-\widehat{\Delta}_b = M$.

It may also affect play in the mechanism, but this change can affect payoffs by at most $(\overline{b} - \underline{s}) + (t_{\max} - t_{\min}) < M$. Hence, the deviation strictly improves the buyer's payoff in these cases. Moreover these cases arise with positive probability (since (b) has probability $(1-\epsilon)/2$). So the deviation is a strict improvement for the buyer.

Consequently, by sequential rationality, the buyer's equilibrium strategy σ_B^* must call for playing A (with probability 1) at each such information set h_B.

Now, given this, consider the informant I_B at stage 1 of the information game. He can make any report with $\widehat{\Delta}_{\underline{b}} = \widehat{\Delta}_{\overline{b}} = -M$, and the buyer will accept, which gives the informant a payoff of ϵM. If instead he chooses inaction, his payoff is 0. So inaction cannot be a best reply.

Similarly, inaction cannot be optimal for I_S given the seller's equilibrium strategy σ_S^*.

Step 2. Suppose that, in the equilibrium outcome, the probability that the buyer rejects, conditional on having value $\underline{b}$, is $\xi_{\underline{b}}$. Then the unconditional probability of rejection is at least $p_{\underline{b}}\xi_{\underline{b}}$. Hence, informant I_B's overall payoff is at most $3M\epsilon - p_{\underline{b}}\xi_{\underline{b}}F$. This must be at least 0 (otherwise the informant would rather deviate to inaction); therefore $\xi_{\underline{b}} \leq 3M\epsilon/p_{\underline{b}}F$. By choice of F, this is at most ϵ^3.

Defining $\xi_{\overline{b}}$ similarly as the buyer's probability of rejecting conditional on value $\overline{b}$, we get $\xi_{\overline{b}} \leq \epsilon^3$.

Similarly, conditional on either value $\underline{s}, \overline{s}$, the seller's probability of rejecting is at most ϵ^3.

From now on, say that a strategy profile σ has the *low-rejection property* if the buyer's probabilities of rejecting conditional on $\underline{b}$ and conditional on $\overline{b}$, and the seller's probabilities of rejecting conditional on $\underline{s}$ and $\overline{s}$, are all $< 2\epsilon^3$.

Step 3. Write $h_B = (\widehat{\underline{k}}(b), \widehat{\Delta}_b, d_B, \eta_B)$ to denote a typical information set of the buyer at the end of the information game. Here, as in the proof of Lemma 4.3, d_B denotes the buyer's decision (A or R) in stage 2. We need not explicitly include the buyer's value b in the description of h_B because b is fully determined by η_B, although sometimes it will be useful to include it. Similarly, we can write $h_S = (\widehat{\underline{k}}(s), \widehat{\Delta}_s, d_S, \eta_S)$ for a typical information set of the seller at the end of the information game.

Let (σ, μ) be any consistent assessment (not necessarily our equilibrium). For any h_B, we can write the implied belief $\mu_B(h_S|h_B)$, a probability distribution over information sets h_S conditional on reaching h_B. This is obtained by taking the probability distribution $\mu_B(z|h_B)$ specified by the belief system μ at information set h_B, which is a distribution over nodes in h_B (all of which are terminal nodes in the information game), and then marginalizing over h_S. In particular, this implied belief is well-defined even if h_B is

reached with probability zero under the proposed strategies.

We will show the following claim: If (σ, μ) is any consistent assessment that satisfies the low-rejection property, then for any two information sets h_B and $h'_B = (\widehat{\underline{k}}'(b), \widehat{\Delta}'_b, d'_B, \eta'_B)$ such that $\eta'_B = \eta_B$, the implied beliefs over h_S are almost the same:

$$d(\mu_B(\cdot|h_B), \mu_B(\cdot|h'_B)) \leq 12\epsilon, \tag{B.1}$$

where the distance $d(\cdot, \cdot)$ between distributions is measured by the L_1-norm.

To see this, first suppose that σ is fully mixed, so that every information set has positive probability and hence μ_B is pinned down by Bayesian updating. Let $\widetilde{\sigma}_S$ be the alternative strategy for the seller that simply always accepts in the information game (play in the mechanism is irrelevant); let $\widetilde{\sigma}$ be the strategy profile where S plays $\widetilde{\sigma}_S$ and the other players B, I_B, I_S play as in σ; and define $\widetilde{\mu}_B(h_S|h_B), \widetilde{\mu}_B(h_S|h'_B)$ to be the implied beliefs over h_S that arise from $\widetilde{\sigma}$. Notice that every information set for B is still reached with positive probability under $\widetilde{\sigma}$, so these alternative beliefs are still pinned down by Bayes.

Now we make the following two subclaims:

(i) For each $h_S = (\widehat{\underline{k}}(s), \widehat{\Delta}_s, d_S, \eta_S)$ with $d_S = A$, we have $\widetilde{\mu}_B(h_S|h_B) \geq (1-2\epsilon^2)\mu_B(h_S|h_B)$.

(ii) The information sets h_S with $d_S = A$ have total probability at least $1 - \epsilon$ under distribution $\mu_B(\cdot|h_B)$.

To see why subclaim (ii) is true, write the desired total probability (slightly abusing notation) as

$$\begin{aligned}
\mu_B(d_S = A|h_B) &= \mu_B(d_S = A|\widehat{\underline{k}}_B, \widehat{\Delta}_B, d_B, \eta_B) \\
&= \frac{\Pr_\sigma(d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \cdot \Pr_\sigma(\eta_B|d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}{\begin{array}{r}\Pr_\sigma(d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \cdot \Pr_\sigma(\eta_B|d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) + \\ \Pr_\sigma(d_S = R; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \cdot \Pr_\sigma(\eta_B|d_S = R; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)\end{array}}.
\end{aligned}$$

Now, in the denominator, the first summand is at least $(1-2\epsilon^3) \times \pi_{\min}\epsilon/(K+1)$, because the probability of $d_S = A$ is at least $1 - 2\epsilon^3$ and this event arises independently of $(b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)$ (since there is no interaction between the buyer's and seller's information or their decisions by the time d_S is decided), and because at stage 3, the signal η_B materializes with probability at least $\epsilon\pi_{\min}/(K+1)$ regardless of all previous events (as one of the possibilities in realization (a)). By our assumption on ϵ, $(1-2\epsilon^3) \times \pi_{\min}\epsilon/(K+1) \geq 2\epsilon^2 \geq 2(\epsilon^2 - \epsilon^3)$.

Likewise, the second summand in the denominator is at most $2\epsilon^3 \times 1 = 2\epsilon^3$, because $d_S = R$ happens with probability at most $2\epsilon^3$ and this is independent of $(d, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)$. Hence, we have

$$\mu_B(d_S = A|h_B) \geq \frac{2(\epsilon^2 - \epsilon^3)}{2(\epsilon^2 - \epsilon^3) + 2\epsilon^3} = 1 - \epsilon.$$

Similarly, for subclaim (i), write

$$\begin{aligned}\widetilde{\mu}_B(h_S|h_B) &= \frac{\Pr_{\widetilde{\sigma}}(h_S, h_B)}{\Pr_{\widetilde{\sigma}}(h_B)} \\ &= \frac{\Pr_{\widetilde{\sigma}}(s, \widehat{\underline{k}}_s, \widehat{\Delta}_s, d_S, b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \times \Pr_{\widetilde{\sigma}}(\eta_S, \eta_B|s, \widehat{\underline{k}}_s, \widehat{\Delta}_s, d_S, b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}{\Pr_{\widetilde{\sigma}}(\eta_B|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \times \Pr_{\widetilde{\sigma}}(b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}.\end{aligned}$$

Here $d_S = A$ by assumption. Now write out a similar formula for $\mu_B(h_S|h_B)$, where all the probabilities are under σ rather than $\widetilde{\sigma}$. Now divide the two formulas. Note that

$$\Pr_{\widetilde{\sigma}}(s, \widehat{\underline{k}}_s, \widehat{\Delta}_s, d_S, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \geq \Pr_{\sigma}(s, \widehat{\underline{k}}_s, \widehat{\Delta}_s, d_S, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)$$

because the seller is more inclined to choose $d_S = A$ under $\widetilde{\sigma}$ than σ, and everything else affecting these probabilities is the same under $\widetilde{\sigma}$ as σ. Also notice that

$$Pr(\eta_S, \eta_B|s, \widehat{\underline{k}}_s, \widehat{\Delta}_s, d_S, b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)$$

is the same under $\widetilde{\sigma}$ as under σ (both probabilities are determined identically by stage 3 of the information game). And $Pr(b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)$ is also clearly the same under $\widetilde{\sigma}$ as σ. So when the smoke clears, we get

$$\frac{\widetilde{\mu}_B(h_S|h_B)}{\mu_B(h_S|h_B)} \geq \frac{\Pr_{\sigma}(\eta_B|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}{\Pr_{\widetilde{\sigma}}(\eta_B|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}.$$

Now the logic is similar to subclaim (ii): break up the numerator and denominator on the right-hand side into

$$\frac{\begin{array}{r}\Pr_{\sigma}(\eta_B|d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \cdot \Pr_{\sigma}(d_S = A|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) + \\ \Pr_{\sigma}(\eta_B|d_S = R; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \cdot \Pr_{\sigma}(d_S = R|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)\end{array}}{\Pr_{\widetilde{\sigma}}(\eta_B|d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \cdot \Pr_{\widetilde{\sigma}}(d_S = A|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}. \tag{B.2}$$

(There is no additional summand in the denominator, since $\widetilde{\sigma}$ places probability zero on R.)

Given that the seller accepts, the probability of producing signal η_B in stage 3 depends only on $\widehat{\underline{k}}_b$ and d_B; in particular, the first factor $Pr(\eta_B|d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)$ is the same under σ and $\widetilde{\sigma}$. So we can divide through by this probability, and also note that the second factor in the denominator is 1, to rewrite (B.2) as

$$\Pr_\sigma(d_S = A|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) + \frac{\Pr_\sigma(\eta_B|d_S = R; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B) \cdot \Pr_\sigma(d_S = R|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}{\Pr_\sigma(\eta_B|d_S = A; b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)}.$$

This is at least $\Pr_\sigma(d_S = A|b, \widehat{\underline{k}}_b, \widehat{\Delta}_b, d_B)$ which is just the seller's probability of accepting (by independence) which is at least $1 - 2\epsilon^2$.

This proves subclaims (i) and (ii). Combining, we see that there is at least $(1-2\epsilon^2)(1-\epsilon) \geq 1 - 3\epsilon$ probability mass that is shared between distributions $\widetilde{\mu}_B(\cdot|h_B)$ and $\mu_B(\cdot|h_B)$. Thus,

$$d(\mu_B(\cdot|h_B), \widetilde{\mu}_B(\cdot|h_B)) \leq 6\epsilon. \tag{B.3}$$

Similarly,

$$d(\mu_B(\cdot|h'_B), \widetilde{\mu}_B(\cdot|h'_B)) \leq 6\epsilon. \tag{B.4}$$

However, we also have

$$\widetilde{\mu}_B(h_S|h_B) = \widetilde{\mu}_B(h_S|h'_B) \tag{B.5}$$

for each information set h_S. This is because $\widetilde{\sigma}_S$ always accepts, so if h_S involves decision $d_S = R$ then both probabilities are zero; on the other hand if $d_S = A$ then we have

$$\begin{aligned} \widetilde{\mu}_B(h_S|h_B) &= \Pr_{\widetilde{\sigma}}(s, \widehat{\underline{k}}_s, \widehat{\Delta}_s) \times \Pr_{\widetilde{\sigma}}(\eta_S|s, \widehat{\underline{k}}_s, \widehat{\Delta}_s, d_S = A; h_B) \\ &= \Pr_{\widetilde{\sigma}}(s, \widehat{\underline{k}}_s, \widehat{\Delta}_s) \times \pi(\eta_S|\eta_B) \end{aligned}$$

since η_B is part of h_B, and once an information structure k is chosen at stage 3, the distribution of η_S given η_B is always $\pi(\eta_S|\eta_B)$ regardless of everything else that has happened. Now, an identical calculation shows that $\widetilde{\mu}_B(h_S|h'_B)$ is given by exactly the same formula, proving (B.5). Finally, combining (B.3), (B.4), and (B.5) (and applying the triangle inequality) gives the claim, inequality (B.1).

This proves (B.1) when the strategy profile σ is fully mixed.

Now if (σ, μ) is any consistent assessment that satisfies the low-rejection property, it is the limit of fully mixed consistent assessments (σ^r, μ^r) with $r \to \infty$. For sufficiently large r, σ^r also satisfies the low-rejection property, so (B.1) holds for each such μ^r, and therefore it holds for the limit μ. This proves the claim in general.

In particular, the claim holds for the equilibrium (σ^*, μ^*), completing this step.

Step 4. Consider the path of play generated by the equilibrium strategies σ^*. This involves some joint distribution over signals (η_B, η_S) and actions in the mechanism (a_B, a_S) (among other things).

For each possible signal η_B, define $\alpha_B(\eta_B)$ to be the probability distribution over a_B conditional on η_B being realized. This is well-defined, because each signal arises with positive probability. Thus $\alpha_B(\eta_B)$ defines a mixed action in the mechanism. Similarly we define $\alpha_S(\eta_S)$ for each possible signal η_S.

Now, just as in Appendix B.1, for each k and each $(\eta_B, \eta_S) \in H_B^k \times H_S^k$, we define $q^k(\eta_B, \eta_S) = q(\alpha_B(\eta_B), \alpha_S(\eta_S))$ and $t^k(\eta_B, \eta_S) = t(\alpha_B(\eta_B), \alpha_S(\eta_S))$, where q and t from the mechanism are extended to mixed actions by linearity.

We will show that these q^k and t^k form an overlapping list of γ-direct mechanisms, where $\gamma = 24M\epsilon$. The overlapping condition (4.3) is satisfied by construction, so we need to check the approximate IC and IR conditions. Fix a signal η_B, let b be the corresponding buyer's value, and let $a_B^* \in A_B$ be an action chosen with positive probability under $\alpha_B(\eta_B)$. For the approximate IC condition, it suffices to show that any deviation from a_B^* to any alternative action a_B' cannot produce a gain of more than γ: that is,

$$\sum_{\eta_S} \pi(\eta_S|\eta_B) \left([bq(a_B^*, \alpha_S(\eta_S)) - t(a_B^*, \alpha_S(\eta_S))] - [bq(a_B', \alpha_S(\eta_S)) - t(a_B', \alpha_S(\eta_S))] \right) \geq -\gamma. \tag{B.6}$$

Indeed, this will imply that there is at most γ gain from deviating from the mixed action $\alpha_B(\eta_B)$ to any a_B', and therefore at most γ gain from deviating to $\alpha_B(\eta_B')$, which is what is needed.

Let h_B be some information set, reached with positive probability, under which the buyer has signal η_B and plays action a_B^* in equilibrium. Write $\Pr(h_S|h_B)$ for the distribution over h_S given h_B (as in step 3; note we can write this as an objective probability distribution rather than a belief μ_B since h_B has positive probability). And write $\Pr(h_S|\eta_B)$ for the equilibrium distribution over h_S conditional only on the buyer receiving signal η_B. Since the latter is an average over the distributions $\Pr(h_S|h_B')$ for various information sets h_B', all with signal η_B, step 3 implies that these two distributions satisfy

$$d_{\Delta(H_S^\cup)}(\Pr(\cdot|h_B), \Pr(\cdot|\eta_B)) \leq 12\epsilon.$$

(The subscript is notation to emphasize that the two $\Pr(\cdots)$'s lie in the space of distributions over $H_S^\cup = \cup_k H_S^k$.) Consequently, the same is true for the distributions over seller's

actions $\sigma_S(h_S)$ conditional on h_B and on η_B respectively:

$$d_{\Delta(A_S)}(\Pr(\cdot|h_B), \Pr(\cdot|\eta_B)) \leq 12\epsilon. \tag{B.7}$$

Now, after conditioning on η_B, the distributions over seller's actions is equal to the weighted average of the distributions obtained by further conditioning on seller's signal η_S, where each η_S occurs with the appropriate probability, which is $\pi(\eta_S|\eta_B)$:

$$\Pr(a_S|\eta_B) = \sum_{\eta_S} \pi(\eta_S|\eta_B) \Pr(a_S|\eta_B, \eta_S). \tag{B.8}$$

However, conditional on the *seller's* information h_S at the time of the mechanism, the choice of a_S is independent of any additional information possessed by the buyer (since the seller cannot condition on this additional information). In particular, a_S is independent of η_B conditional on h_S:

$$\Pr(a_S, \eta_B|h_S) = \Pr(a_S|h_S) \Pr(\eta_B|h_S) = \pi(\eta_B|\eta_S) \Pr(a_S|h_S).$$

Doing a weighted sum over all information sets h_S at which the seller receives signal η_S gives

$$\begin{aligned} \Pr(a_S, \eta_B|\eta_S) &= \sum_{h_S} \Pr(a_S, \eta_B|h_S) \Pr(h_S|\eta_S) \\ &= \pi(\eta_B|\eta_S) \sum_{h_S} \Pr(a_S|h_S) \Pr(h_S|\eta_S) \\ &= \pi(\eta_B|\eta_S) \Pr(a_S|\eta_S) \end{aligned}$$

or, after dividing by the probability of η_B given η_S,

$$\Pr(a_S|\eta_B, \eta_S) = \Pr(a_S|\eta_S).$$

Thus (B.8) becomes

$$\Pr(a_S|\eta_B) = \sum_{\eta_S} \pi(\eta_S|\eta_B) \Pr(a_S|\eta_S). \tag{B.9}$$

Combining with (B.7) gives a bound on the distance between the distribution of the seller's actions conditional on the buyer being at h_B, and the weighted average of the

distributions $\alpha_S(\eta_S)$:

$$d_{\Delta(A_S)}(\Pr(\cdot|h_B), \sum_{\eta_S} \pi(\eta_S|\eta_B)\alpha_S(\eta_S)) \leq 12\epsilon. \tag{B.10}$$

Now, let $\widehat{\alpha}_S(h_B)$ denote the distribution over actions a_S conditional on the buyer's information set h_B. Because a change in the seller's action affects the buyer's payoff in the mechanism by at most M, (B.10) implies

$$\left| [bq(a_B^*, \widehat{\alpha}_S(h_B)) - t(a_B^*, \widehat{\alpha}_S(h_B))] - \sum_{\eta_S} \pi(\eta_S|\eta_B) \left[bq(a_B^*, \alpha_S(\eta_S)) - t(a_B^*, \alpha_S(\eta_S))\right] \right| \leq 12M\epsilon \tag{B.11}$$

and

$$\left| [bq(a_B', \widehat{\alpha}_S(h_B)) - t(a_B', \widehat{\alpha}_S(h_B))] - \sum_{\eta_S} \pi(\eta_S|\eta_B) \left[bq(a_B', \alpha_S(\eta_S)) - t(a_B', \alpha_S(\eta_S))\right] \right| \leq 12M\epsilon. \tag{B.12}$$

And since the equilibrium strategy prescribes playing a_B^* at h_B, sequential rationality gives

$$[bq(a_B^*, \widehat{\alpha}_S(h_B)) - t(a_B^*, \widehat{\alpha}_S(h_B))] - [bq(a_B', \widehat{\alpha}_S(h_B)) - t(a_B', \widehat{\alpha}_S(h_B))] \geq 0.$$

Combining with (B.11) and (B.12) (and the triangle inequality) gives (B.6). This implies that the approximate IC condition for the buyer is satisfied. And by taking a_B' to be the non-participation action $\emptyset$ in (B.6), we get the approximate IR condition as well.

The corresponding conditions for the seller hold by analogous reasoning.

Step 5. Fix any signal η_B that the buyer may receive, and let $b \in \{\underline{b}, \overline{b}\}$ be the associated value. Let $u_B^*(\eta_B)$ denote the buyer's expected payoff, conditional on receiving signal η_B, in the direct mechanism constructed in Step 4; that is,

$$u_B^*(\eta_B) = \sum_{\eta_S} \pi(\eta_S|\eta_B) \left[bq(\alpha_B(\eta_B), \alpha_S(\eta_S)) - t(\alpha_B(\eta_B), \alpha_S(\eta_S))\right].$$

(Note that this is the correct payoff formula for any direct mechanism (q^k, t^k) for an information structure in which η_B appears.)

We claim the following: Under the equilibrium assessment (σ^*, μ^*), starting from any information structure h_B at which the buyer receives signal η_B, her expected payoff is within $36M\epsilon$ of $u_B^*(\eta_B)$.

To show this, we need more notation. First, let $\widetilde{\alpha}_S(\eta_B)$ denote the equilibrium distribution over seller's actions in the mechanism conditional on the buyer receiving signal η_B. From (B.9), we have

$$\widetilde{\alpha}_S(\eta_B) = \sum_{\eta_S} \pi(\eta_S|\eta_B)\alpha_S(\eta_S).$$

Also let $\widehat{\alpha}_S(h_B)$ be the distribution over $\sigma_S(h_S)$ when h_S is distributed according to $\mu_B(h_S|h_B)$, just as in Step 4.

The same reasoning leading to (B.7) implies that for any action a_B that B could take in the mechanism,

$$|[bq(a_B, \widehat{\alpha}_S(h_B)) - t(a_B, \widehat{\alpha}_S(h_B))] - [bq(a_B, \widetilde{\alpha}_S(\eta_B)) - t(a_B, \widetilde{\alpha}_S(\eta_B))]| \leq 12M\epsilon. \quad \text{(B.13)}$$

Here we again use the fact that a change in the seller's action in the mechanism can affect the buyer's payoff by at most M. (In Step 4, we assumed that h_B was an information set reached with positive probability, but that property is not needed here.)

Now, if a_B^* is any action in the support of $\alpha_B(\eta_B)$, then it is an action taken with positive probability at some information set h_B' with signal η_B. Hence, by (B.6), it gives a payoff that is within $\gamma = 24M\epsilon$ of optimal against action distribution $\widetilde{\alpha}_S(\eta_B)$ by the seller. Consequently, the mixed action $\alpha_B(\eta_B)$ is also within γ of optimal against $\widetilde{\alpha}_S(\eta_B)$. That is, $u_B^*(\eta_B)$ is within $24M\epsilon$ of the best payoff the buyer can get against $\widetilde{\alpha}_S(\eta_B)$. By (B.13), the latter is within $12M\epsilon$ of the best payoff the buyer can get against $\widehat{\alpha}_S(h_B)$. But this latter quantity is exactly the payoff that the buyer gets starting from h_B. This proves the claim.

Now, for each $b \in \{\underline{b}, \overline{b}\}$, write

$$u_b(\mathcal{M}^k) = \sum_{\eta_B} \pi^k(\eta_B|b)u_B^*(\eta_B)$$

for the expected payoff of the buyer in direct mechanism $\mathcal{M}^k$ given that her value is b, exactly as with the definitions of $u_{\underline{b}}, u_{\overline{b}}$ in Subsection 4.1.

Consider now any information set $(b, \widehat{\underline{k}}_b, \widehat{\Delta}_b)$ at which the buyer has value b and has to choose to accept or reject extortion. For each $k = 0, \ldots, K$, let ψ^k be the probability, as assessed at this information set, that one of the two below cases occurs: either

- stage 3 of the information game will lead to realization (a) with information structure $\mathcal{S}^k$ being chosen; or

- realization (c) will occur, and $\underline{\widehat{k}}(s) = k$, and the seller chose R.

In either of these cases, no matter whether the buyer chooses A or R, information structure k will be realized and $g_B = 0$. Let $\overline{\psi}$ be the probability that either (b) or (c) realizes and the seller chose A; in this case, the buyer choosing A leads to information structure 0 and $g_B = -\widehat{\Delta}_b$, whereas R leads to $k = \underline{\widehat{k}}(b)$ and $g_B = 0$. Finally, let $\widetilde{\psi}^k$ be the probability that (b) realizes, $\underline{\widehat{k}}(s) = k$, and the seller chose R; in this case the buyer's A/R action chooses between information structure k and $g_B = -\widehat{\Delta}_b$, and $\underline{\widehat{k}}(b)$ and $g_B = 0$.

So if the buyer chooses A at the current information set $(b, \underline{\widehat{k}}_b, \widehat{\Delta}_b)$, her expected payoff in the combined game is within $36M\epsilon$ of

$$\sum_k \psi^k u_b(\mathcal{M}^k) + \overline{\psi}(u_b(\mathcal{M}^0) - \widehat{\Delta}_b) + \sum_k \widetilde{\psi}^k(u_b(\mathcal{M}^k) - \widehat{\Delta}_b), \tag{B.14}$$

and if she chooses R, her expected payoff in the combined game is within $36M\epsilon$ of

$$\sum_k \psi^k u_b(\mathcal{M}^k) + \overline{\psi}(u_b(\mathcal{M}^{\underline{\widehat{k}}_b})) + \sum_k \widetilde{\psi}^k(u_b(\mathcal{M}^{\underline{\widehat{k}}_b})). \tag{B.15}$$

Subtracting: the difference in payoffs from choosing A versus choosing R is within $72M\epsilon$ of

$$\overline{\psi}(u_b(\mathcal{M}^0) - u_b(\mathcal{M}^{\underline{\widehat{k}}_b}) - \widehat{\Delta}_b) + \sum_k \widetilde{\psi}^k(u_b(\mathcal{M}^k) - u_b(\mathcal{M}^{\underline{\widehat{k}}_b}) - \widehat{\Delta}_b). \tag{B.16}$$

Now, we know from step 2 that the seller's probability of rejecting is $< 2\epsilon^3$, and this is independent of the information $(b, \underline{\widehat{k}}_b, \widehat{\Delta}_b)$ received so far. Therefore, $\sum_k \widetilde{\psi}^k < 2\epsilon^3$ and $\overline{\psi} \geq ((1-\epsilon)/2) \times (1 - 2\epsilon^3) \geq 1/4$ (by choice of ϵ). Combining with the fact that $|u_b(\mathcal{M}^k) - u_b(\mathcal{M}^{\underline{\widehat{k}}_b})| \leq M$ and $|\widehat{\Delta}_b| \leq M$, we see that the difference in payoffs between choosing A and R is within $76M\epsilon$ of

$$\overline{\psi}(u_b(\mathcal{M}^0) - u_b(\mathcal{M}^{\underline{\widehat{k}}(b)}) - \widehat{\Delta}_b).$$

So if

$$u_b(\mathcal{M}^0) - u_b(\mathcal{M}^{\underline{\widehat{k}}(b)}) - \widehat{\Delta}_b > 400M\epsilon, \tag{B.17}$$

then this choosing A is strictly better than choosing R.

Hence, the buyer's equilibrium strategy must call for choosing A at every information set $(b, \underline{\widehat{k}}_b, \widehat{\Delta}_b)$ where (B.17) is satisfied.

Now define the true worst information structure $\underline{k}(b)$ and true willingness-to-pay Δ_b, for each buyer type b, as in Appendix B.1. It follows that in equilibrium, the informant

I_B must report $\widehat{\Delta}_b \geq \Delta_b - 400M\epsilon - \delta$ (for each buyer type b) with probability 1: If this were not the case for some buyer type b, then the informant could increase his report $\widehat{\Delta}_b$ while having it still be below $\Delta_b - 400M\epsilon$ and could report $\widehat{\underline{k}}(b) = \underline{k}(b)$, and this would increase the informant's payoff, since the buyer of type b would still accept this higher extortion. (Note the δ term that comes from the discretization of the informant's action space.)

By a similar argument, in equilibrium, the informant I_S must with probability 1 make a report that satisfies $\widehat{\Delta}_s \geq \Delta_s - 400M\epsilon - \delta$, for each type of seller $s \in \{\underline{s}, \overline{s}\}$.

Step 6. Finally, we can estimate the ex-ante expected equilibrium payoff of each player.

We know that with probability at least $(1-2\epsilon^3)^2(1-\epsilon) \geq 1-5\epsilon$, both players choose A and then either (b) or (c) is realized in stage 3 of the information game. When this happens, the resulting information structure will be $k=0$ and the players' payoffs in the information game are given by $g_B = -\widehat{\Delta}_b, g_S = -\widehat{\Delta}_s$. In the remaining 5ϵ of probability mass, these players' information-game payoffs will in any case be different than the above by at most $2M$ (since payoffs in the information game are always in $[-M, M]$). Hence, the buyer's expected payoff in the information game is

$$\leq -\mathbb{E}[\widehat{\Delta}_b] + 10M\epsilon \leq -\mathbb{E}[\Delta_b] + 410M\epsilon + \delta$$

where the expectations are with respect to $b \in \{\underline{b}, \overline{b}\}$.

What about the payoff in the mechanism? Step 5 also showed us that, conditional on receiving any signal η_B, the buyer's expected payoff in the mechanism is within $36M\epsilon$ of $u_B^*(\eta_B)$; hence, buyer type b's expected payoff conditional on information structure $k=0$ realizing is within $36M\epsilon$ of $u_b(\mathcal{M}^0)$. Again, information structure 0 realizes with probability at least $1-5\epsilon$, and in the remaining 5ϵ mass of cases, the buyer's payoff in the mechanism cannot differ by more than M. So the buyer's ex-ante expected payoff in the mechanism is

$$\leq \mathbb{E}[u_b(\mathcal{M}^0)] + 41M\epsilon.$$

Hence, the buyer's ex-ante expected payoff in the combined game is

$$\leq \mathbb{E}[u_b(\mathcal{M}^0) - \Delta_b] + 451M\epsilon + \delta = \mathbb{E}[u_b(\mathcal{M}^{\underline{k}(b)})] + 451M\epsilon + \delta.$$

A similar bound applies to the seller's ex-ante expected payoff. Finally, each informant can never obtain a payoff of more than $3M\epsilon$.

Thus, adding up across the buyer, seller, and two informants, total equilibrium welfare

in the combined game is at most

$$\mathbb{E}[u_b(\mathcal{M}^{\underline{k}(b)})] + \mathbb{E}[u_s(\mathcal{M}^{\underline{k}(s)})] + 908M\epsilon + 2\delta = TMU(\mathcal{L}) + 908M\epsilon + 2\delta$$

where $\mathcal{L}$ is the list of γ-direct mechanisms constructed in step 4, and its total minimum utility $TMU(\mathcal{L})$ is defined as in (4.2).

Wrapping up the proof. At this point we have shown the following: For suitably small ϵ, δ (with M/δ an integer), for any sequential equilibrium of the corresponding combined game, total equilibrium utility is at most $TMU(\mathcal{L}_\gamma) + 908M\epsilon + 2\delta$, for some $\mathcal{L}_\gamma$ that is an overlapping list of γ-direct mechanisms with $\gamma = 24M\epsilon$.

Now, different choices of the equilibrium may lead to different choices of mechanism list $\mathcal{L}_\gamma$. Nonetheless, for all such choices, welfare satisfies an upper bound of $\max_{\mathcal{L}_\gamma} TMU(\mathcal{L}_\gamma) + 908M\epsilon + 2\delta$, where the max is over all overlapping lists of γ-direct mechanisms. By taking $\mathcal{L}^*_\gamma$ to be a list that attains the max (it is straightforward to check that this exists), we then get the bound

$$\overline{W}(\mathcal{M}) \leq \overline{W}(\mathcal{M}, \mathcal{I}) \leq TMU(\mathcal{L}^*_\gamma) + 908M\epsilon + 2\delta.$$

Now take $\epsilon \to 0$ and $\delta \to 0$. By compactness arguments, we can assume that the direct mechanisms $\mathcal{L}^*_\gamma$ converge to some limiting $\mathcal{L}^*$. It is straightforward to check that $\mathcal{L}^*$ is an overlapping list of (0-)direct mechanisms. We thus get

$$\overline{W}(\mathcal{M}) \leq TMU(\mathcal{L}^*)$$

and the lemma follows. □

C Computations for upper bounds on welfare

Here we give the details on the upper bounds on $TMU(\mathcal{L})$, for each region in Section 5.

In the proofs of Lemmas 5.1 and 5.2, since the states have been labeled by value pairs, we can represent a direct mechanism by functions $q : \{\underline{b}, \overline{b}\} \times \{\underline{s}, \overline{s}\} \to [0, 1]$ and $t : \{\underline{b}, \overline{b}\} \times \{\underline{s}, \overline{s}\} \to \mathbb{R}$ as in Subsection 2.2.[8] We also adopt the notation $u_B(b, s), u_S(b, s)$ for the players' payoffs, as in that subsection.

[8]This representation may lose information, since it does not describe the outcome specified by the mechanism when the players report a signal pair that can never occur. However, this will not be a problem for proving the lemmas.

Proof of Lemma 5.1. Let $\mathcal{L} = (\mathcal{M}^1, \mathcal{M}^2)$ be a list of direct mechanisms for the given $(\mathcal{S}^1, \mathcal{S}^2)$ in Table 3. We will show that

$$\begin{aligned} p_{\overline{b}} u_{\overline{b}}(\mathcal{M}^1) + p_{\overline{s}} u_{\overline{s}}(\mathcal{M}^1) &\leq p_{\overline{b}}(\overline{b} - \overline{s}), & \text{(C.1)} \\ p_{\underline{b}} u_{\underline{b}}(\mathcal{M}^2) + p_{\underline{s}} u_{\underline{s}}(\mathcal{M}^2) &\leq p_{\underline{s}}(\underline{b} - \underline{s}). & \text{(C.2)} \end{aligned}$$

We will just prove (C.1) here; the proof of (C.2) is symmetrical. First, in state $(\underline{b}, \overline{s})$, both players know the state and trade is not efficient, so the two players' IR constraints imply that both get payoff zero in this state. Then

$$\begin{aligned} p_{\overline{b}} u_{\overline{b}}(\mathcal{M}^1) + p_{\overline{s}} u_{\overline{s}}(\mathcal{M}^1) &= p_{\overline{b}} p_{\underline{s}} u_B(\overline{b}, \underline{s}) + p_{\overline{b}} p_{\overline{s}} u_B(\overline{b}, \overline{s}) + p_{\overline{b}} p_{\overline{s}} u_S(\overline{b}, \overline{s}) \\ &= p_{\overline{b}} p_{\underline{s}} u_B(\overline{b}, \underline{s}) + p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) q(\overline{b}, \overline{s}). & \text{(C.3)} \end{aligned}$$

The IC constraint for the seller in state $(\overline{b}, \underline{s})$ (who can pretend to have value $\overline{s}$) implies $u_S(\overline{b}, \underline{s}) \geq u_S(\overline{b}, \overline{s}) + (\overline{s} - \underline{s}) q(\overline{b}, \overline{s})$. Since the two parties' payoffs add up to at most the total surplus $b - s$, we have

$$\begin{aligned} u_B(\overline{b}, \underline{s}) &\leq (\overline{b} - \underline{s}) - \left[u_S(\overline{b}, \overline{s}) + (\overline{s} - \underline{s}) q(\overline{b}, \overline{s}) \right] \\ &\leq (\overline{b} - \underline{s}) - (\overline{s} - \underline{s}) q(\overline{b}, \overline{s}) \end{aligned}$$

since $u_S(\overline{b}, \overline{s}) \geq 0$ by the seller's IR.

Therefore,

$$\begin{aligned} p_{\overline{b}} p_{\underline{s}} u_B(\overline{b}, \underline{s}) + p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) q(\overline{b}, \overline{s}) &\leq p_{\overline{b}} p_{\underline{s}} \left[(\overline{b} - \underline{s}) - (\overline{s} - \underline{s}) q(\overline{b}, \overline{s}) \right] + p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) q(\overline{b}, \overline{s}) \\ &= p_{\overline{b}} \left(p_{\underline{s}} (\overline{b} - \underline{s}) + \left(p_{\overline{s}} (\overline{b} - \overline{s}) - p_{\underline{s}} (\overline{s} - \underline{s}) \right) q(\overline{b}, \overline{s}) \right) & \text{(C.4)} \end{aligned}$$

The assumption $p_{\underline{s}} \leq p^*_{\underline{s}}$ implies $p_{\overline{s}} (\overline{b} - \overline{s}) - p_{\underline{s}} (\overline{s} - \underline{s}) \geq 0$, and therefore the right side of (C.4) is maximized over $q(\overline{b}, \overline{s}) \in [0, 1]$ by taking $q(\overline{b}, \overline{s}) = 1$. Hence we get

$$\begin{aligned} p_{\overline{b}} p_{\underline{s}} u_B(\overline{b}, \underline{s}) + p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) q(\overline{b}, \overline{s}) &\leq p_{\overline{b}} \left(p_{\underline{s}} (\overline{b} - \underline{s}) + p_{\overline{s}} (\overline{b} - \overline{s}) - p_{\underline{s}} (\overline{s} - \underline{s}) \right) \\ &= p_{\overline{b}} \left(p_{\underline{s}} (\overline{b} - \overline{s}) + p_{\overline{s}} (\overline{b} - \overline{s}) \right) \\ &= p_{\overline{b}} (\overline{b} - \overline{s}). \end{aligned}$$

Combining with (C.3) gives (C.1), as needed.

Now putting together (C.1) and (C.2) gives the result. □

Proof of Lemma 5.2. Let $\mathcal{L} = (\mathcal{M}^1, \mathcal{M}^2)$ be a list of direct mechanisms. Note that our proof of (C.1) in region I used only the assumption $p_{\underline{s}} \leq p^*_{\underline{s}}$, which is still valid here, and never used $p_{\overline{b}} \leq p^*_{\overline{b}}$. Thus (C.1) still holds.

Let $\lambda = \frac{p_{\underline{b}}}{p_{\overline{b}}} \times \frac{\overline{s}-\underline{s}}{\overline{b}-\underline{b}}$. The parameter assumption $p_{\overline{b}} > p^*_{\overline{b}}$ implies $\lambda \in (0,1)$.

We will show that any mechanism $\mathcal{M}^2$ for $\mathcal{S}^2$ satisfies

$$\begin{aligned} p_{\underline{b}} u_{\underline{b}}(\mathcal{M}^2) + (1-\lambda) p_{\overline{b}} u_{\overline{b}}(\mathcal{M}^2) + p_{\underline{s}} u_{\underline{s}}(\mathcal{M}^2) + (1-\lambda) p_{\overline{s}} u_{\overline{s}}(\mathcal{M}^2) \\ \leq p_{\overline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) + (1-\lambda) p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}). \end{aligned} \tag{C.5}$$

To begin the argument, note that in state $(\underline{b}, \overline{s})$, both players know the state and IR ensures both players get payoff zero. In state $(\overline{b}, \overline{s})$, the two players' payoffs add up to at most the maximum feasible surplus $\overline{b} - \overline{s}$.

Now, when the seller has type $\underline{s}$, she does not know the buyer's value, so the IC constraint for the buyer in state $(\overline{b}, \underline{s})$, and then the IR in $(\underline{b}, \underline{s})$, give

$$u_B(\overline{b}, \underline{s}) \geq u_B(\underline{b}, \underline{s}) + (\overline{b} - \underline{b}) q(\underline{b}, \underline{s}) \geq (\overline{b} - \underline{b}) q(\underline{b}, \underline{s}).$$

Hence

$$\begin{aligned} p_{\underline{b}} u_B(\underline{b}, \underline{s}) + (1-\lambda) p_{\overline{b}} u_B(\overline{b}, \underline{s}) + u_S(\underline{s}) \\ &= p_{\underline{b}} (u_B(\underline{b}, \underline{s}) + u_S(\underline{b}, \underline{s})) + p_{\overline{b}} (u_B(\overline{b}, \underline{s}) + u_S(\overline{b}, \underline{s})) - \lambda p_{\overline{b}} u_B(\overline{b}, \underline{s}) \\ &\leq p_{\underline{b}} (\underline{b} - \underline{s}) q(\underline{b}, \underline{s}) + p_{\overline{b}} (\overline{b} - \underline{s}) q(\overline{b}, \underline{s}) - \lambda p_{\overline{b}} (\overline{b} - \underline{b}) q(\underline{b}, \underline{s}) \\ &= \left[p_{\underline{b}} (\underline{b} - \underline{s}) - \lambda p_{\overline{b}} (\overline{b} - \underline{b}) \right] q(\underline{b}, \underline{s}) + p_{\overline{b}} (\overline{b} - \underline{s}) q(\overline{b}, \underline{s}) \\ &\leq p_{\overline{b}} (\overline{b} - \underline{s}) \end{aligned} \tag{C.6}$$

(where we have used, in the third line, that in each state (b, s), the two players' payoffs add up to the total surplus $(b-s)q(b,s)$; and in going from the fourth to the fifth line, the fact that the bracketed expression is zero by definition of λ).

Thus, the left side of (C.5) can be expanded to

$$p_{\underline{b}} p_{\underline{s}} u_B(\underline{b}, \underline{s}) + (1-\lambda) p_{\overline{b}} p_{\underline{s}} u_B(\overline{b}, \underline{s}) + (1-\lambda) p_{\overline{b}} p_{\overline{s}} u_B(\overline{b}, \overline{s}) + p_{\underline{s}} p_{\underline{b}} u_S(\underline{b}, \underline{s}) + p_{\underline{s}} p_{\overline{b}} u_S(\overline{b}, \underline{s}) + (1-\lambda) p_{\overline{s}} p_{\overline{b}} u_S(\overline{b}, \overline{s})$$

(where we have already removed the terms associated with state $(\underline{b}, \overline{s})$, which are zero)

$$\begin{aligned} &\leq p_{\underline{s}} \left(p_{\underline{b}} u_B(\underline{b}, \underline{s}) + (1-\lambda) p_{\overline{b}} u_B(\overline{b}, \underline{s}) + p_{\underline{b}} u_S(\underline{b}, \underline{s}) + p_{\overline{b}} u_S(\overline{b}, \underline{s}) \right) \\ &\qquad + (1-\lambda) p_{\overline{b}} p_{\overline{s}} \left(u_B(\overline{b}, \overline{s}) + u_S(\overline{b}, \overline{s}) \right) \\ &\leq p_{\underline{s}} p_{\overline{b}} (\overline{b} - \underline{s}) + (1-\lambda) p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) \end{aligned}$$

by (C.6) and the fact that total surplus under $(\overline{b}, \overline{s})$ is at most $\overline{b} - \overline{s}$. This proves (C.5).

Now, adding (C.5) to λ times (C.1) gives the bound:

$$\begin{aligned} TMU(\mathcal{L}) &\leq \left(p_{\underline{b}} u_{\underline{b}}(\mathcal{M}^2) + (1-\lambda) p_{\overline{b}} u_{\overline{b}}(\mathcal{M}^2) + p_{\underline{s}} u_{\underline{s}}(\mathcal{M}^2) + (1-\lambda) p_{\overline{s}} u_{\overline{s}}(\mathcal{M}^2) \right) \\ &\qquad + \lambda \left(p_{\overline{b}} u_{\overline{b}}(\mathcal{M}^1) + p_{\overline{s}} u_{\overline{s}}(\mathcal{M}^1) \right) \\ &\leq p_{\overline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) + (1-\lambda) p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) + \lambda p_{\overline{b}} (\overline{b} - \overline{s}) \\ &= p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) + p_{\overline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) + \lambda p_{\overline{b}} (1 - p_{\overline{s}}) (\overline{b} - \overline{s}) \\ &= p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) + p_{\overline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) + \left(p_{\underline{b}} \frac{\underline{b} - \underline{s}}{\overline{b} - \underline{b}} \right) \times p_{\underline{s}} (\overline{b} - \overline{s}) \end{aligned}$$

which is the desired bound.

□

For parameter region III, recall that a new information structure $\mathcal{S}^2$ was introduced, with its states labeled by pairs $(b, s) \in \{\underline{b}, \overline{b}, \overline{b}'\} \times \{\underline{s}, \overline{s}\}$. We will again represent a mechanism by functions q, t defined on such pairs.

Proof of Lemma 5.4. As noted in the text, we have $\lambda \in (0, 1)$. Also note that our parameter assumption implies

$$\lambda \leq \frac{p_{\overline{s}}}{p_{\underline{s}}} \times \frac{\overline{b} - \overline{s}}{\overline{s} - \underline{s}}. \tag{C.7}$$

Let $\mathcal{L} = (\mathcal{M}^1, \mathcal{M}^2)$ be an overlapping list of direct mechanisms. We will show the following bounds:

$$\begin{aligned} p_{\overline{b}} u_{\overline{b}}(\mathcal{M}^1) + p_{\overline{s}} u_{\overline{s}}(\mathcal{M}^1) + (1-\lambda) p_{\overline{b}} p_{\underline{s}} u_S^1(\overline{b}, \underline{s}) &\leq p_{\overline{b}} p_{\overline{s}} (\overline{b} - \overline{s}) + p_{\overline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) \\ &\qquad - p_{\underline{b}} p_{\underline{s}} (\overline{s} - \underline{s}) \frac{\underline{b} - \underline{s}}{\overline{b} - \underline{b}}, \end{aligned} \tag{C.8}$$

$$p_{\underline{b}} u_{\underline{b}}(\mathcal{M}^2) + p_{\underline{b}} p_{\underline{s}} u_S^2(\underline{b}, \underline{s}) + \lambda p_{\overline{b}} p_{\underline{s}} u_S^2(\overline{b}, \underline{s}) \leq p_{\underline{b}} p_{\underline{s}} (\overline{b} - \underline{s}) \frac{\underline{b} - \underline{s}}{\overline{b} - \underline{b}}. \tag{C.9}$$

Here the superscripts on u_S^1, u_S^2 refer to the mechanism (and information structure) in

which the payoff is being evaluated.

To prove (C.8), consider mechanism $\mathcal{M}^1$ for information structure $\mathcal{S}^1$. The IC for seller type $\underline{s}$ in state $(\overline{b}, \underline{s})$, and IR for the seller in state $(\overline{b}, \overline{s})$, give

$$u_S(\overline{b}, \underline{s}) \geq u_S(\overline{b}, \overline{s}) + (\overline{s} - \underline{s})q(\overline{b}, \overline{s}) \geq (\overline{s} - \underline{s})q(\overline{b}, \overline{s}).$$

Taking into account that both parties get payoff zero in state $(\underline{b}, \overline{s})$, the left-hand side of (C.8) is

$$\begin{aligned}
&p_{\overline{b}}p_{\underline{s}}u_B(\overline{b}, \underline{s}) + p_{\overline{b}}p_{\overline{s}}u_B(\overline{b}, \overline{s}) + p_{\overline{b}}p_{\overline{s}}u_S(\overline{b}, \overline{s}) + (1-\lambda)p_{\overline{b}}p_{\underline{s}}u_S(\overline{b}, \underline{s}) \\
&\qquad = p_{\overline{b}}\left(p_{\underline{s}}\left(u_B(\overline{b}, \underline{s}) + u_S(\overline{b}, \underline{s})\right) + p_{\overline{s}}\left(u_B(\overline{b}, \overline{s}) + u_S(\overline{b}, \overline{s})\right) - \lambda p_{\underline{s}}u_S(\overline{b}, \underline{s})\right) \\
&\qquad = p_{\overline{b}}\left(p_{\underline{s}}(\overline{b} - \underline{s})q(\overline{b}, \underline{s}) + p_{\overline{s}}(\overline{b} - \overline{s})q(\overline{b}, \overline{s}) - \lambda p_{\underline{s}}u_S(\overline{b}, \underline{s})\right) \\
&\qquad \leq p_{\overline{b}}\left(p_{\underline{s}}(\overline{b} - \underline{s})q(\overline{b}, \underline{s}) + p_{\overline{s}}(\overline{b} - \overline{s})q(\overline{b}, \overline{s}) - \lambda p_{\underline{s}}(\overline{s} - \underline{s})q(\overline{b}, \overline{s})\right) \\
&\qquad = p_{\overline{b}}\left(p_{\underline{s}}(\overline{b} - \underline{s})q(\overline{b}, \underline{s}) + \left[p_{\overline{s}}(\overline{b} - \overline{s}) - \lambda p_{\underline{s}}(\overline{s} - \underline{s})\right]q(\overline{b}, \overline{s})\right)
\end{aligned}$$

Because the bracketed expression is ≥ 0 by (C.7), and because $q(\overline{b}, \underline{s}), q(\overline{b}, \overline{s}) \leq 1$, we get

$$\leq \quad p_{\overline{b}}\left(p_{\underline{s}}(\overline{b} - \underline{s}) + \left[p_{\overline{s}}(\overline{b} - \overline{s}) - \lambda p_{\underline{s}}(\overline{s} - \underline{s})\right]\right)$$

which is the right-hand side of (C.8). Thus (C.8) holds.

To prove (C.9), consider mechanism $\mathcal{M}^2$. As usual, because B gets payoff 0 in state $(\underline{b}, \overline{s})$, the left side simplifies to

$$p_{\underline{s}}\left(p_{\underline{b}}u_B(\underline{b}, \underline{s}) + p_{\underline{b}}u_S(\underline{b}, \underline{s}) + \lambda p_{\overline{b}}u_S(\overline{b}, \underline{s})\right).$$

Using the IC of type $\overline{b}$ in state $(\overline{b}, \underline{s})$ and IR of type $\underline{b}$ in state $(\underline{b}, \underline{s})$, we have

$$u_B(\overline{b}, \underline{s}) \geq u_B(\underline{b}, \underline{s}) + (\overline{b} - \underline{b})q(\underline{b}, \underline{s}) \geq (\overline{b} - \underline{b})q(\underline{b}, \underline{s})$$

and therefore

$$u_S(\overline{b}, \underline{s}) \leq (\overline{b} - \underline{s}) - (\overline{b} - \underline{b})q(\underline{b}, \underline{s}).$$

So

$$\begin{aligned} p_{\underline{b}}u_B(\underline{b},\underline{s}) + p_{\underline{b}}u_S(\underline{b},\underline{s}) + \lambda p_{\overline{b}}u_S(\overline{b},\underline{s}) &\leq p_{\underline{b}}(\underline{b}-\underline{s})q(\underline{b},\underline{s}) + \lambda p_{\overline{b}}\left[(\overline{b}-\underline{s}) - (\overline{b}-\underline{b})q(\underline{b},\underline{s})\right] \\ &= \left[p_{\underline{b}}(\underline{b}-\underline{s}) - \lambda p_{\overline{b}}(\overline{b}-\underline{b})\right] q(\underline{b},\underline{s}) + \lambda p_{\overline{b}}(\overline{b}-\underline{s}) \\ &= \lambda p_{\overline{b}}(\overline{b}-\underline{s}) \end{aligned}$$

since the bracketed factor is zero. Multiplying through by $p_{\underline{s}}$, and plugging in the value of λ, gives (C.9).

Now, note that

$$\begin{aligned} u_{\underline{s}}(\mathcal{M}^2) &= p_{\underline{b}}u_S^2(\underline{b},\underline{s}) + \lambda p_{\overline{b}}u_S^2(\overline{b},\underline{s}) + (1-\lambda)p_{\overline{b}}u_S^2(\overline{b}',\underline{s}) \\ &= p_{\underline{b}}u_S^2(\underline{b},\underline{s}) + \lambda p_{\overline{b}}u_S^2(\overline{b},\underline{s}) + (1-\lambda)p_{\overline{b}}u_S^1(\overline{b},\underline{s}) \end{aligned}$$

because $u_S^1(\overline{b},\underline{s}) = u_S^2(\overline{b}',\underline{s})$ by the overlapping condition. Thus, when we add together the left sides of (C.8) and (C.9), the three $\underline{s}$ terms combine into $p_{\underline{s}}u_{\underline{s}}(\mathcal{M}^2)$. Thus, adding together the two inequalities gives

$$p_{\underline{b}}u_{\underline{b}}(\mathcal{M}^2) + p_{\overline{b}}u_{\overline{b}}(\mathcal{M}^1) + p_{\underline{s}}u_{\underline{s}}(\mathcal{M}^2) + p_{\overline{s}}u_{\overline{s}}(\mathcal{M}^1) \leq p_{\overline{b}}p_{\overline{s}}(\overline{b}-\overline{s}) + p_{\overline{b}}p_{\underline{s}}(\overline{b}-\underline{s}) + p_{\underline{b}}p_{\underline{s}}(\overline{b}-\overline{s})\frac{\underline{b}-\underline{s}}{\overline{b}-\underline{b}}$$

which is exactly what we need.

□

D Additional omitted proofs

Proof of Lemma 2.1. It is straightforward to check that the mechanisms shown are dominant-strategy incentive-compatible. (In particular, note that in the left mechanism, when the buyer's type is $\overline{b}$, the seller of type $\underline{s}$ gets a payoff of $\underline{b}-\underline{s}$ from either report; similarly for the buyer of type $\overline{b}$ in the mechanism on the right.) So we just need to show that no mechanism can exceed the welfare formulas stated.

As above, for any dominant-strategy mechanism, incentive-compatibility implies $u_B(\overline{b},\underline{s}) \geq u_B(\underline{b},\underline{s}) + (\overline{b}-\underline{b})q(\underline{b},\underline{s}) \geq (\overline{b}-\underline{b})q(\underline{b},\underline{s})$, and similarly $u_S(\overline{b},\underline{s}) \geq (\overline{s}-\underline{s})q(\overline{b},\overline{s})$. Adding gives

$$(\overline{b}-\underline{b})q(\underline{b},\underline{s}) + (\overline{s}-\underline{s})q(\overline{b},\overline{s}) \leq u_B(\overline{b},\underline{s}) + u_S(\overline{b},\underline{s}) = (\overline{b}-\underline{s})q(\overline{b},\underline{s}) \leq \overline{b}-\underline{s}.$$

Let us rewrite this as

$$(\overline{b}-\underline{b})(1-q(\underline{b},\underline{s}))+(\overline{s}-\underline{s})(1-q(\overline{b},\overline{s}))\geq \overline{s}-\underline{b}. \tag{D.1}$$

Now, in case (a), consider the welfare shortfall relative to first-best; it must satisfy

$$\begin{aligned}
&\left[p_{\underline{b}}p_{\underline{s}}(\underline{b}-\underline{s})+p_{\overline{b}}p_{\underline{s}}(\overline{b}-\underline{s})+p_{\overline{b}}p_{\overline{s}}(\overline{b}-\overline{s})\right]-W\\
&\qquad\geq p_{\underline{b}}p_{\underline{s}}(\underline{b}-\underline{s})(1-q(\underline{b},\underline{s}))+p_{\overline{b}}p_{\overline{s}}(1-q(\overline{b},\overline{s}))\\
&\qquad\geq p_{\overline{b}}p_{\overline{s}}\frac{\overline{b}-\underline{b}}{\overline{s}-\underline{s}}(\overline{b}-\overline{s})(1-q(\underline{b},\underline{s}))+p_{\overline{b}}p_{\overline{s}}(\overline{b}-\overline{s})(1-q(\overline{b},\overline{s}))\\
&\qquad\geq p_{\overline{b}}p_{\overline{s}}\frac{\overline{b}-\overline{s}}{\overline{s}-\underline{s}}(\overline{s}-\underline{b})
\end{aligned}$$

where the first inequality is by $q(\overline{b},\underline{s})\leq 1$, the second is by the assumption of case (a), and the third is by (D.1). Rearranging shows that W is bounded above by the expression stated in (a), as needed.

The proof of case (b) is symmetric.

□

Proof of Proposition 2.2. First note that any dominant-strategy mechanism certainly satisfies the constraints of the Bayesian problem. Now suppose we are in case (a) of Lemma 2.1 (case (b) is analogous).

Consider the optimal dominant-strategy mechanism, on the left side of Table 1. Notice that the Bayesian incentive constraint for the buyer of type $\underline{b}$ to report type $\overline{b}$ is satisfied with strict inequality (since misreporting has no effect if the seller is type $\underline{s}$, and only hurts the buyer if the seller is type $\overline{s}$). Likewise, the Bayesian incentive constraint for the seller of type $\underline{s}$ to report type $\overline{s}$ is satisfied as a strict inequality (the seller is indifferent if the buyer is $\overline{b}$, and is strictly hurt by misreporting if $\underline{b}$).

Now change $q(\overline{b},\overline{s})$ from $\frac{b-s}{\overline{s}-\underline{s}}$ to $\frac{b-s}{\overline{s}-\underline{s}}+\epsilon$, and change $t(\overline{b},\overline{s})$ from $\frac{b-s}{\overline{s}-\underline{s}}\,\overline{s}$ to $\left(\frac{b-s}{\overline{s}-\underline{s}}+\epsilon\right)\overline{s}$, where $\epsilon>0$. If ϵ is small enough, this change cannot lead to a violation of the Bayesian IC constraints of types $\underline{b},\underline{s}$, since these constraints were originally slack; and it also cannot violate the other IC constraints or the IR constraints since it only makes types $\overline{b},\overline{s}$ better off. So the new mechanism still satisfies all constraints, and evidently has a strictly higher welfare than before. □

Proof of Proposition 3.1. Assume the parameters are as in case (a) of Lemma 2.1. (Case

(b) is analogous.) If $p_{\underline{s}} \leq p^*_{\underline{s}}$, then the condition for case (a) can be rearranged to give

$$\frac{p_{\underline{b}}}{p_{\overline{b}}} \geq \frac{p_{\overline{s}}}{p_{\underline{s}}} \times \frac{\overline{b}-\overline{s}}{\overline{s}-\underline{s}} \times \frac{\overline{b}-\underline{b}}{\underline{b}-\underline{s}} \geq \frac{\overline{b}-\underline{b}}{\underline{b}-\underline{s}}$$

from which $p_{\overline{b}} \leq p^*_{\underline{b}}$ also. Moreover, if $p_{\underline{s}} < p^*_{\underline{s}}$ then $p_{\overline{b}} < \underline{b}^*$ strictly, by the same logic.

Now consider the difference $W_{FP} - W_{DS}$. By subtracting $p_{\underline{b}}p_{\underline{s}}(\underline{b}-\underline{s})$ from both W_{FP} and W_{DS}, we compute

$$\begin{aligned}
W_{FP} - W_{DS} &= \left[p_{\overline{b}}p_{\underline{s}}(\underline{b}-\underline{s}) + p_{\overline{b}}(\overline{b}-\overline{s})\right] - \left[p_{\overline{b}}p_{\underline{s}}(\overline{b}-\underline{s}) + p_{\overline{b}}p_{\overline{s}}\frac{(\overline{b}-\overline{s})(\underline{b}-\underline{s})}{\overline{s}-\underline{s}}\right] \\
&= p_{\overline{b}}\left(\left[p_{\underline{s}}(\underline{b}-\underline{s}) + (p_{\underline{s}}+p_{\overline{s}})(\overline{b}-\overline{s})\right] - \left[p_{\underline{s}}(\overline{b}-\underline{s}) + p_{\overline{s}}\frac{(\overline{b}-\overline{s})(\underline{b}-\underline{s})}{\overline{s}-\underline{s}}\right]\right) \\
&= p_{\overline{b}}\left(p_{\underline{s}}(\underline{b}-\overline{s}) + p_{\overline{s}}(\overline{b}-\overline{s}) - p_{\overline{s}}\frac{(\overline{b}-\overline{s})(\underline{b}-\underline{s})}{\overline{s}-\underline{s}}\right) \\
&= p_{\overline{b}}(\overline{s}-\underline{b})\left(-p_{\underline{s}} + p_{\overline{s}}\frac{\overline{b}-\overline{s}}{\overline{s}-\underline{s}}\right).
\end{aligned}$$

So evidently $W_{FP} - W_{DS}$ is the same sign as $-p_{\underline{s}}/p_{\overline{s}} + (\overline{b}-\overline{s})/(\overline{s}-\underline{s})$; the latter expression is zero exactly at $p_{\underline{s}} = p^*_{\underline{s}}$.

Thus:

- If $p_{\underline{s}} < p^*_{\underline{s}}$ (and then also $p_{\overline{b}} < p^*_{\overline{b}}$), then $W_{FP} - W_{DS} > 0$.
- If $p_{\underline{s}} = p^*_{\underline{s}}$ (and then $p_{\overline{b}} \leq \overline{b}^*$), then $W_{FP} - W_{DS} = 0$.
- If $p_{\underline{s}} > p^*_{\underline{s}}$, then $W_{FP} - W_{DS} < 0$.

So in all three subcases, the comparison between W_{FP} and W_{DS} runs as claimed in the proposition. □

Proof of Lemma B.1. Consider the buyer, in the auxiliary game, after receiving some signal η^*_B. If she plays according to her proposed equilibrium action $\alpha_B(\eta^*_B)$, her expected payoff is

$$\sum_{\eta_S} \pi(\eta_S|\eta^*_B)\left[b(\eta^*_B)q(\alpha_B(\eta^*_B), \alpha_S(\eta_S)) - t(\alpha_B(\eta_B), \alpha^*_S(\eta_S))\right],$$

where the sum is over all η_S lying in any H^k_S. (Here the conditional probability $\pi(\eta_S|\eta^*_B)$

is unambiguously defined, as described in Subsection 4.1.) We can rewrite this as

$$\sum_{\eta_S \in H_S^k} \pi(\eta_S|\eta_B^*) \left[b(\eta_B^*) q^k(\eta_B^*, \eta_S) - t^k(\eta_B^*, \eta_S) \right] \tag{D.2}$$

where k is any index with $\eta_B^* \in H_B^k$. If instead she deviates to $\alpha_B(\eta_B')$ for some η_B', then her expected payoff is likewise

$$\sum_{\eta_S \in H_S^k} \pi(\eta_S|\eta_B^*) \left[b(\eta_B^*) q^k(\eta_B', \eta_S) - t^k(\eta_B', \eta_S) \right]. \tag{D.3}$$

The equilibrium condition says that (D.2) must be greater than or equal to (D.3). Multiplying through by $\pi^k(\eta_B^*)$, and re-expressing both sides in terms of states in Ω^k, we obtain the IC constraint for the buyer in a direct mechanism on $\mathcal{S}^k$.

Also, the the buyer can always get a nonnegative payoff by playing $\emptyset$ in the mechanism, equilibrium implies that (D.2) must be at least 0. Multiplying by $\pi^k(\eta_B^*)$ and re-expressing in terms of states gives us the buyer IR constraint for a direct mechanism.

Similarly, we see that the seller's IC and IR constraints are all satisfied.

And the overlapping condition (4.3) is trivially satisfied, by the construction of the q^k and t^k. □

Proof of Theorem 3.2. First, we prove the upper bound on $\underline{W}(\mathcal{M})$ in each parameter region, for any (indirect) mechanism $\mathcal{M}$. (We can restrict to information games without additional players, or allow additional players; the arguments are the same.)

If $p_{\overline{b}} \leq p_{\overline{b}}^*$ and $p_{\underline{s}} \leq p_{\underline{s}}^*$ (region I), then apply the extortion lemma, Lemma 4.3, to the information structures in Table 3. The lemma gives us $\underline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L})$. Combining with Lemma 5.1 gives $\underline{W}(\mathcal{M}) \leq p_{\overline{b}}(\overline{b} - \overline{s}) + p_{\underline{s}}(\underline{b} - \underline{s}) = W_{FP}$.

Now suppose $p_{\overline{b}} > p_{\overline{b}}^*$ but $p_{\underline{s}} \leq p_{\underline{s}}^*$ (region II). Then likewise combining Lemma 4.3 with Lemma 5.2 gives $\underline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}) \leq W_{DS}$ where W_{DS} is determined by case (b) of Lemma 2.1. If $p_{\underline{s}} > p_{\underline{s}}^*$ but $p_{\overline{b}} \leq p_{\overline{b}}^*$ (region II′) then we combine Lemmas 4.3 and 5.3 to obtain $\underline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}) \leq W_{DS}$, where now W_{DS} is given by case (a) of Lemma 2.1.

This leaves us with the case where $p_{\overline{b}} > p_{\overline{b}}^*$ and $p_{\underline{s}} > p_{\underline{s}}^*$. In this case, if $p_{\underline{b}} p_{\underline{s}} \frac{\underline{b}-\underline{s}}{\overline{b}-\underline{b}} \leq p_{\overline{b}} p_{\overline{s}} \frac{\overline{b}-\overline{s}}{\overline{s}-\underline{s}}$, we are in region III; combining Lemmas 4.3 and 5.4 gives $\underline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}) \leq W_{DS}$ in this case. Otherwise, we are in region III′, and combining Lemmas 4.3 and 5.5 gives $\underline{W}(\mathcal{M}) \leq \max_{\mathcal{L}} TMU(\mathcal{L}) \leq W_{DS}$.

This shows the upper bound in all cases. Next we need to show that the upper bound is

attained (again, it does not matter whether we allow additional players in the information game). As mentioned in the text, we cannot quite use the dominant-strategy and flexible-price mechanisms as originally formulated because of the worst-case equilibrium selection criterion. We instead slightly modify them as follows.

First consider the dominant-strategy mechanism on the left side of Table 1. We modify it to a new mechanism $\mathcal{M}'$ as follows: $A_B = \{\emptyset, -, \underline{b}, \overline{b}\}$, $A_S = \{\emptyset, -, \underline{s}, \overline{s}\}$, and the q and t functions are as shown on the left side of Table 5. Here x is an arbitrary positive number, and we have explicitly included the non-participation action for clarity.

For buyer type $\underline{b}$, action $\underline{b}$ now weakly dominates all other actions. For buyer type $\overline{b}$, action $\overline{b}$ weakly dominates all other actions. Similarly for the seller: $\underline{s}$ weakly dominates all other actions for type $\underline{s}$, and $\overline{s}$ weakly dominates all other actions for type $\overline{s}$. Thus, no matter what the information game $\mathcal{I}$ is, in any undominated sequential equilibrium, each agent reports her type truthfully at the mechanism stage, both on and off the equilibrium path. Then each player's equilibrium payoff must be at least her payoff from truthful play of the mechanism (since she has the option of being inactive in the information game and then reporting truthfully in the mechanism), hence total welfare in the combined game is at least the total welfare in the mechanism, which is W_{DS}. A symmetric argument applies when the optimal dominant-strategy mechanism is the one on the right side of Table 1.

For the flexible-price mechanism, we do the same construction. Instead of the mechanism shown in Table 2, which represents a flexible-price mechanism with the seller offering, we take the version on the right side of Table 5. For type $\underline{b}$ of buyer, action $\underline{b}$ weakly dominates all others; for type $\overline{b}$, action $\overline{s}$ weakly dominates all others. For type $\underline{s}$ of the seller, both actions $\underline{b}$ and $\overline{s}$ are undominated, but $\emptyset$ and $-$ are dominated by $\underline{b}$; for type $\overline{s}$, action $\overline{s}$ dominates all others. So in any undominated equilibrium, types $\underline{b}, \overline{b}, \overline{s}$ always play their unique undominated action in the mechanism, and $\underline{s}$ plays either $\underline{b}$ or $\overline{s}$. In particular, in the combined game, the buyer of type $\overline{b}$ always has the option of being inactive in the information game and then playing $\overline{b}$ in the mechanism, which assures her a payoff at least $\overline{b} - \overline{s}$; and the seller of type $\underline{s}$ has the option of being inactive in the information game and then playing $\underline{s}$ in the mechanism, which assures her $\underline{b} - \underline{s}$. And types $\underline{b}, \overline{s}$ are assured at least zero by non-participation. So, any undominated sequential equilibrium of the combined game gives a total welfare at least $p_{\overline{b}}(\overline{b} - \overline{s}) + p_{\underline{s}}(\underline{b} - \underline{s}) = W_{FP}$.

This shows that, for all parameters, both W_{DS} and W_{FP} are indeed attainable values of $\underline{W}(\mathcal{M})$, as needed.

□

$-$	$q:0$ $t:0$	$q:0$ $t:-x$	$q:0$ $t:-x$	$q:0$ $t:0$
$\overline{s}$	$q:0$ $t:0$	$q:0$ $t:0$	$q:\frac{\overline{b}-\underline{s}}{\overline{s}-\underline{s}}$ $t:\frac{\overline{b}-\underline{s}}{\overline{s}-\underline{s}}\,\overline{s}$	$q:0$ $t:x$
$\underline{s}$	$q:0$ $t:0$	$q:1$ $t:\underline{b}$	$q:1$ $t:\underline{b}$	$q:0$ $t:x$
$\emptyset$	$q:0$ $t:0$	$q:0$ $t:0$	$q:0$ $t:0$	$q:0$ $t:0$
	$\emptyset$	$\underline{b}$	$\overline{b}$	$-$

$-$	$q:0$ $t:0$	$q:0$ $t:-x$	$q:0$ $t:-x$	$q:0$ $t:0$
$\overline{s}$	$q:0$ $t:0$	$q:0$ $t:0$	$q:1$ $t:\overline{s}$	$q:0$ $t:x$
$\underline{b}$	$q:0$ $t:0$	$q:1$ $t:\underline{b}$	$q:1$ $t:\underline{b}$	$q:0$ $t:x$
$\emptyset$	$q:0$ $t:0$	$q:0$ $t:0$	$q:0$ $t:0$	$q:0$ $t:0$
	$\emptyset$	$\underline{b}$	$\overline{s}$	$-$

Table 5: Modified versions of dominant-strategy (left) and flexible-price (right) mechanisms. Columns are buyer's actions, rows are seller's. $x > 0$ is arbitrary.

Proof of Theorem 3.3. The proof of the upper bound on $\overline{W}(\mathcal{M})$ for any mechanism $\mathcal{M}$ is exactly as for Theorem 3.2, using Lemma 4.4 instead of Lemma 4.3.

It remains to show that the bound is attained: in particular, that the dominant-strategy mechanism from Lemma 2.1 attains the welfare guarantee W_{DS}, and the flexible-price mechanism attains the welfare guarantee W_{FP}, in some sequential equilibrium of the combined game (for any $\mathcal{I}$). This was already argued in the text. More precisely, consider the combined game, but suppose the players' actions are restricted by never allowing them to play the action $\emptyset$ in the mechanism. The argument in Subsection 3.2 shows that the resulting game has an equilibrium that achieves welfare W_{DS} or W_{FP}, respectively. This remains an equilibrium when we allow action $\emptyset$, since it is never strictly preferred over the existing actions. □

Proof of Proposition 6.1. We first describe equilibrium strategies in the mechanism. The buyer of type $\underline{b}$ accepts if offered price $\underline{b}$, and rejects price $\overline{s}$. The buyer of type $\overline{b}$ accepts both prices. Clearly this is a (weakly) dominant strategy for each buyer type.

The seller, if type $\overline{s}$, offers trade at price $\overline{s}$ (either deterministic or probabilistic, depending on whether branch (i) or (ii) of the mechanism realizes). If the seller's type is $\underline{s}$, her choice of what to do depends on her posterior belief about the buyer's type, which comes from her signal:

- If branch (i) is realized (probability $1-\delta$), the seller offers price $\underline{b}$ if she places posterior probability at most $(\underline{b}-\underline{s})/(\overline{s}-\underline{s})$ on the high type of buyer $\overline{b}$, and otherwise she offers price $\overline{s}$.

- If branch (ii) is realized (probability δ), the seller offers price $\underline{b}$ if she places posterior probability at most $(1/q_\epsilon) \times (\underline{b} - \underline{s})/(\overline{s} - \underline{s})$ on the high type of buyer $\overline{b}$, and otherwise she offers (probabilistic trade at) price $\overline{s}$.

Given the buyer's strategy above, this is optimal for the seller. Hence, the proposed strategies do form an equilibrium. We now need to show that the resulting expected welfare is bounded strictly above W_{FP}, regardless of the information structure $\mathcal{S}$.

First, consider any signal η_S that the seller may receive if her value is $\underline{s}$. Let p be the probability that the buyer has value $\overline{b}$ conditional on the seller receiving η_S. We will show that the expected welfare conditional on the seller receiving signal η_S satisfies

$$\mathbb{E}[\text{welfare} \mid \eta_S] \geq p[(\overline{b} - \underline{s}) + \delta(\overline{s} - \underline{b})] + (\underline{b} - \underline{s}) - p(\overline{s} - \underline{s}). \tag{D.4}$$

We show this in three cases:

- $p \leq (\underline{b} - \underline{s})/(\overline{s} - \underline{s})$. Then, in both branches, the seller offers price $\underline{b}$, and trade takes place with certainty. Hence, expected welfare is $p(\overline{b} - \underline{s}) + (1 - p)(\underline{b} - \underline{s})$. We check that this satisfies (D.4): this is equivalent to

  $$(1 - p)(\underline{b} - \underline{s}) \geq p\delta(\overline{s} - \underline{b}) + (\underline{b} - \underline{s}) - p(\overline{s} - \underline{s})$$

  or

  $$0 \geq p[\delta(\overline{s} - \underline{b}) + (\underline{b} - \underline{s}) - (\overline{s} - \underline{s})] = p(\delta - 1)(\overline{s} - \underline{b})$$

  which is true.

- $\frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}} < p \leq \frac{1}{q_\epsilon} \times \frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}}$. Then in branch (i) the seller offers price $\overline{s}$, but in branch (ii) she offers $\underline{b}$. Then, if the buyer actually has value $\overline{b}$, trade occurs (deterministically) in both branches. If the buyer has value $\underline{b}$, then no trade occurs in the first branch, and trade occurs (deterministically) in the second branch. Hence, expected welfare is

  $$p(\overline{b} - \underline{s}) + (1 - p)\delta(\underline{b} - \underline{s}).$$

  This satisfies (D.4) if and only if

  $$(1 - p)\delta(\underline{b} - \underline{s}) \geq p\delta(\overline{s} - \underline{b}) + (\underline{b} - \underline{s}) - p(\overline{s} - \underline{s}). \tag{D.5}$$

  Since both sides are linear in p, it suffices to check that (D.5) is satisfied at both ends of the interval $[\frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}}, 1]$. At the lower endpoint $p = \frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}}$, both sides of (D.5) are

equal. At the upper endpoint $p = 1$, the inequality reduces to $0 \geq (\delta - 1)(\overline{s} - \underline{b})$, which is true.

- $p > \frac{1}{q_\epsilon} \times \frac{\underline{b} - \underline{s}}{\overline{s} - \underline{s}}$. Then, the seller offers price $\overline{s}$ (with trade being deterministic in branch (i), probabilistic in branch (ii)), which is accepted if and only if the buyer has value $\overline{b}$. Hence, if trade is realized, it produces surplus of $\overline{b} - \underline{s}$; so expected welfare in this case is

$$p((1 - \delta) + \delta q_\epsilon)(\overline{b} - \underline{s}).$$

Thus (D.4) is equivalent to

$$p\delta(-1 + q_\epsilon)(\overline{b} - \underline{s}) \geq p\delta(\overline{s} - \underline{b}) + (\underline{b} - \underline{s}) - p(\overline{s} - \underline{s}). \tag{D.6}$$

Now, when $\delta = 0$, the difference between the left side and right side of (D.6) is $p(\overline{s} - \underline{s}) - (\underline{b} - \underline{s}) \geq \left(\frac{1}{q_\epsilon} - 1\right)(\underline{b} - \underline{s}) > 0$. It follows that for any sufficiently small δ (given fixed ϵ), the difference remains positive for all p, so (D.4) continues to hold.

Thus, as long as δ is chosen small enough, (D.4) holds for each η_S such that the seller's value is $\underline{s}$.

Now we can give a lower bound on expected welfare conditional *only* on the seller having value $\underline{s}$, by taking expectations of (D.4) over all η_S. Note that the right side is linear in p, and on average p must equal the prior probability of the high buyer value, $p_{\overline{b}}$. Thus,

$$\begin{aligned} \mathbb{E}[\text{welfare} \mid \underline{s}] &\geq p_{\overline{b}}[(\overline{b} - \underline{s}) + \delta(\overline{s} - \underline{b})] + (\underline{b} - \underline{s}) - p_{\overline{b}}(\overline{s} - \underline{s}) \\ &= p_{\overline{b}}[(\overline{b} - \overline{s}) + \delta(\overline{s} - \underline{b})] + (\underline{b} - \underline{s}). \end{aligned} \tag{D.7}$$

What about expected welfare when the seller has value $\overline{s}$? Since the seller only offers price $\overline{s}$, which is accepted only by the $\overline{b}$ buyer, welfare in this case is

$$\mathbb{E}[\text{welfare} \mid \overline{s}] = p_{\overline{b}}((1 - \delta) + \delta q_\epsilon)(\overline{b} - \overline{s}).$$

Combining the two cases, overall expected welfare is at least

$$p_{\underline{s}}\left[p_{\overline{b}}[(\overline{b} - \overline{s}) + \delta(\overline{s} - \underline{b})] + (\underline{b} - \underline{s})\right] + p_{\overline{s}}p_{\overline{b}}(1 - \delta + \delta q_\epsilon)(\overline{b} - \overline{s}). \tag{D.8}$$

It remains to check that this expression is strictly higher than W_{FP}. We may view

it as a linear function of δ. When $\delta = 0$, it equals $p_{\overline{b}}(\overline{b} - \overline{s}) + p_{\underline{s}}(\underline{b} - \underline{s}) = W_{FP}$. The derivative with respect to δ is

$$p_{\underline{s}}p_{\overline{b}}(\overline{s} - \underline{b}) + p_{\overline{s}}p_{\overline{b}}(-1 + q_\epsilon)(\overline{b} - \overline{s}) = p_{\overline{b}}p_{\overline{s}}\left((-1 + q_\epsilon) + \frac{p_{\underline{s}}}{p_{\overline{s}}} \times \frac{\overline{s} - \underline{b}}{\overline{b} - \overline{s}}\right) \times (\overline{b} - \overline{s}).$$

The choice of ϵ ensures that the middle factor is positive, so the entire expression is positive. So for $\delta > 0$, the expression in (D.8) is strictly above W_{FP}. Since we have shown that the equilibrium welfare of the mechanism is bounded below by (D.8) regardless of the information structure $\mathcal{S}$, the proposition follows.

□

E Counterexamples for simpler information structures

We detail here the counterexamples mentioned in Section 6.3, showing that various attempts to simplify the proof of the main theorem would not succeed.

E.1 Assigning types to information structures

We show that in region II, we cannot pre-assign types to information structures as in region I. That is, we show that for some parameters in this region, there exists a list of direct mechanisms $(\mathcal{M}^1, \mathcal{M}^2)$ for the information structures $(\mathcal{S}^1, \mathcal{S}^2)$ shown in Table 3, such that

$$p_{\underline{b}}u_{\underline{b}}(\mathcal{M}^2) + p_{\overline{b}}u_{\overline{b}}(\mathcal{M}^1) + p_{\underline{s}}u_{\underline{s}}(\mathcal{M}^2) + p_{\overline{s}}u_{\overline{s}}(\mathcal{M}^1) > W_{DS}.$$

Specifically, we take $(\underline{s}, \underline{b}, \overline{s}, \overline{b}) = (1, 2, 3, 4)$, so that $p^*_{\overline{b}} = p^*_{\underline{s}} = 1/3$. Then take $p_{\overline{b}} = 1/2$ and $p_{\underline{s}} = 1/4$. These parameters indeed lie in region II. The dominant-strategy welfare bound given there is easily computed to be $13/16$. However, for each of the two information structures, we can consider a posted-price mechanism:

- for $\mathcal{S}^1$, a posted price of 3 (accepted by both seller types and by the high-value buyer);
- for $\mathcal{S}^2$, a posted price of 4 (again accepted by both seller types and the high-value buyer).

Then

$$p_{\underline{b}}u_{\underline{b}}(\mathcal{M}^2)+p_{\overline{b}}u_{\overline{b}}(\mathcal{M}^1)+p_{\underline{s}}u_{\underline{s}}(\mathcal{M}^2)+p_{\overline{s}}u_{\overline{s}}(\mathcal{M}^1) = \frac{1}{2}\times 0+\frac{1}{2}\times 1+\frac{1}{4}\times 2+\frac{3}{4}\times 0 = 1 > \frac{13}{16} = W_{DS}.$$

$\overline{s}$	$u_B = 0$ $u_S = 0$	$u_B = 6/7$ $u_S = 0$
$\underline{s}$	$u_B = 0$ $u_S = 1$	$u_B = 9/7$ $u_S = 12/7$
	$\underline{b}$	$\overline{b}$

$\overline{s}$	$u_B = 0$ $u_S = 0$	$u_B = 1$ $u_S = 0$
$\underline{s}$	$u_B = 0$ $u_S = 4/7$	$u_B = 8/7$ $u_S = 13/7$
	$\underline{b}$	$\overline{b}$

Table 6: Payoffs from example mechanisms for information structures from regions I, II, applied in region III.

E.2 Information structures in region III

Here we show that the list of information structures $(\mathcal{S}^1, \mathcal{S}^2)$ used in regions I, II would not suffice for a tight bound in region III. We give an example of parameters in this region and a list $\mathcal{L}$ of direct mechanisms such that $TMU(\mathcal{L}) > W_{DS}$ strictly.

We again take $(\underline{s}, \underline{b}, \overline{s}, \overline{b}) = (1, 2, 3, 4)$, and $p_{\overline{b}} = 3/4$, $p_{\underline{s}} = 1/2$. These parameters indeed lie in region III. The optimal dominant-strategy welfare is computed to be $W_{DS} = 25/16$. We consider the following two mechanisms $\mathcal{M}^1, \mathcal{M}^2$ for information structures $\mathcal{S}^1, \mathcal{S}^2$ respectively:

- For $\mathcal{S}^1$, where the seller is fully informed: If the buyer is type $\overline{b}$ (uninformed), the seller can choose either to sell with probability 6/7, at a price of 3 conditional on trade; or to sell for sure at a price of 19/7. (We can check that the seller of type $\underline{s}$ takes the second option, and the seller of type $\overline{s}$ takes the first.) If the buyer is type $\underline{b}$, the seller can sell at a price of 2 (only type $\underline{s}$ takes this up).

- For $\mathcal{S}^2$, where the buyer is fully informed: If the seller is type $\underline{s}$ (uninformed), the buyer has the choice of either buying with probability 4/7, at a price of 2 conditional on trade; or buying for sure at a price of 20/7. (Check that the buyer of type $\underline{b}$ takes the first option, and type $\overline{b}$ takes the second.) If the seller is type $\overline{s}$, then the buyer can buy at a price of 3 (only type $\overline{b}$ takes this up).

The resulting payoff of each agent, for each possible pair of values (b, s), is as shown in Table 6.

We can check that each type has the same expected payoff in $\mathcal{M}^1$ as in $\mathcal{M}^2$. Types $\underline{b}$ and $\overline{s}$ have payoff zero in both; and

$$u_{\overline{b}}(\mathcal{M}^1) = u_{\overline{b}}(\mathcal{M}^2) = 15/14, \qquad u_{\underline{s}}(\mathcal{M}^1) = u_{\underline{s}}(\mathcal{M}^2) = 43/28.$$

Therefore,

$$\begin{aligned} TMU(\mathcal{L}) &= p_{\underline{b}} u_{\underline{b}}(\mathcal{L}) + p_{\overline{b}} u_{\overline{b}}(\mathcal{L}) + p_{\underline{s}} u_{\underline{s}}(\mathcal{L}) + p_{\overline{s}} u_{\overline{s}}(\mathcal{L}) \\ &= \frac{1}{4} \times 0 + \frac{3}{4} \times \frac{15}{14} + \frac{1}{2} \times \frac{43}{28} + \frac{1}{2} \times 0 \\ &= \frac{11}{7} \\ &> \frac{25}{16} \quad = W_{DS}. \end{aligned}$$

E.3 Overlapping information structures

Here we show that the argument for region III could not have used the non-overlapping version of the extortion lemma. We give an example with a non-overlapping list $\mathcal{L}$ of direct mechanisms (on the same information structures) where $TMU(\mathcal{L}) > W_{DS}$.

We again take $(\underline{s}, \underline{b}, \overline{s}, \overline{b}) = (1, 2, 3, 4)$, $p_{\overline{b}} = 3/4$, $p_{\underline{s}} = 1/2$, so that $W_{DS} = 25/16$. The information structures are those shown in Table 4, where $\lambda = 1/6$.

In describing information structure $\mathcal{S}^2$, we will refer to $\underline{b}, \overline{b}, \overline{b}'$ (in the labeling of the states) as "quasi-types" of the buyer, to distinguish them from the types (values) which are only $\underline{b}, \overline{b}$.

We consider the following mechanisms for each information structure:

- For $\mathcal{S}^1$, where the seller is fully informed: If the buyer is type $\overline{b}$ (uninformed), the seller can either sell with probability 8/9, at a price of 13/4 conditional on trade; or to sell for sure at a price 3. (The seller of type $\underline{s}$ is willing to take the second option, and $\overline{s}$ takes the first.) If the buyer is type $\underline{b}$, the seller can sell at a price of 1 (only type $\underline{s}$ takes this up.)

  (Since there is full information about the buyer's value, we can indeed design the mechanism on the $\overline{b}$ states independently from the $\underline{b}$ states, without worrying about incentive-compatibility for the buyer.)

- For $\mathcal{S}^2$: If the seller is type $\underline{s}$ and the buyer is not quasi-type $\overline{b}'$ (so the seller does not know whether the buyer is $\underline{b}$ or $\overline{b}$), then the good is sold at price 1. If the seller is $\overline{s}$ and the buyer is not $\overline{b}'$, then the seller can offer the good at price 3 (and only quasi-type $\overline{b}$ of buyer takes the sale). Finally, when the buyer is $\overline{b}'$ (and so is uninformed of the seller's value), then the good is sold at price 17/5.

The payoff of each agent in each state is as shown in Table 7.

$\overline{s}$	$u_B = 0$ $u_S = 0$	$u_B = 2/3$ $u_S = 2/9$
$\underline{s}$	$u_B = 1$ $u_S = 0$	$u_B = 1$ $u_S = 2$
	$\underline{b}$	$\overline{b}$

$\overline{s}$	$u_B = 0$ $u_S = 0$	$u_B = 1$ $u_S = 0$	$u_B = 3/5$ $u_S = 2/5$
$\underline{s}$	$u_B = 1$ $u_S = 0$	$u_B = 3$ $u_S = 0$	$u_B = 3/5$ $u_S = 12/5$
	$\underline{b}$	$\overline{b}$	$\overline{b}'$

Table 7: Payoffs in example of non-overlapping mechanisms in region III.

These are indeed non-overlapping mechanisms: the outcome in $\mathcal{M}^1$ when the buyer has high value $\overline{b}$ is different than the outcome in $\mathcal{M}^2$ when the buyer has quasi-type $\overline{b}'$.

We can again check the payoff of each of the types $\underline{b}, \overline{b}, \underline{s}, \overline{s}$ in each mechanism. Note that in information structure 2, the buyer is quasi-type $\overline{b}$ with marginal probability $\lambda p_{\overline{b}} = 1/8$ and $\overline{b}'$ with probability $(1-\lambda)p_{\overline{b}} = 5/8$, and $u_{\overline{b}}(\mathcal{M}^2)$ is computed by combining these cases.

$$u_{\underline{b}}(\mathcal{M}^1) = u_{\underline{b}}(\mathcal{M}^2) = 1/2, \qquad u_{\overline{b}}(\mathcal{M}^1) = u_{\overline{b}}(\mathcal{M}^2) = 5/6,$$
$$u_{\underline{s}}(\mathcal{M}^1) = u_{\underline{s}}(\mathcal{M}^2) = 3/2, \qquad u_{\overline{s}}(\mathcal{M}^1) = 1/6, \qquad u_{\overline{s}}(\mathcal{M}^2) = 1/4.$$

Therefore, the total minimum utility is

$$\begin{aligned} TMU(\mathcal{L}) &= p_{\underline{b}} u_{\underline{b}}(\mathcal{L}) + p_{\overline{b}} u_{\overline{b}}(\mathcal{L}) + p_{\underline{s}} u_{\underline{s}}(\mathcal{L}) + p_{\overline{s}} u_{\overline{s}}(\mathcal{L}) \\ &= \frac{1}{4} \times \frac{1}{2} + \frac{3}{4} \times \frac{5}{6} + \frac{1}{2} \times \frac{3}{2} + \frac{1}{2} \times \frac{1}{6} \\ &= \frac{19}{12} \\ &> \frac{25}{16} \quad = W_{DS}. \end{aligned}$$

References

[1] Dirk Bergemann and Stephen Morris (2005), "Robust Mechanism Design," *Econometrica* 73 (6), 1771–1813.

[2] Dirk Bergemann, Benjamin Brooks, and Stephen Morris (2016), "Informationally Robust Optimal Auction Design," unpublished paper, University of Chicago.

[3] Dirk Bergemann and Juuso Välimäki (2002), "Information Acquisition and Efficient Mechanism Design," *Econometrica* 70 (3), 1007–1033.

[4] Sushil Bikhchandani (2010), "Information Acquisition and Full Surplus Extraction," *Journal of Economic Theory* 145 (6), 2282–2308.

[5] Tilman Börgers (2017), "(No) Foundations of Dominant-Strategy Mechanisms: A Comment on Chung and Ely (2007)," *Review of Economic Design* 21 (2), 73–82.

[6] Tilman Börgers and Doug Smith (2012), "Robustly Ranking Mechanisms," *American Economic Review* 102 (3), 325–329.

[7] Benjamin Brooks and Songzi Du (2019), "Optimal Auction Design with Common Values: An Informationally-Robust Approach," unpublished paper, University of Chicago.

[8] Gabriel Carroll and Ilya Segal (2018), "Robustly Optimal Auctions with Unknown Resale Opportunities," *Review of Economic Studies*, forthcoming.

[9] Kim-Sau Chung and J. C. Ely (2007), "Foundations of Dominant-Strategy Mechanisms," *Review of Economic Studies* 74 (2), 447–476.

[10] Jacques Crémer and Fahad Khalil (1992), "Gathering Information Before Signing a Contract," *American Economic Review* 82 (3), 566–578.

[11] Jacques Crémer, Fahad Khalil, and Jean-Charles Rochet (1998), "Contracts and Productive Information Gathering," *Games and Economic Behavior* 25 (2), 174–193.

[12] Jacques Crémer, Fahad Khalil, and Jean-Charles Rochet (1998), "Strategic Information Gathering Before a Contract is Offered," *Journal of Economic Theory* 81 (1), 163–200.

[13] Drew Fudenberg and Jean Tirole (1991), *Game Theory*, Cambridge, Massachusetts: MIT Press.

[14] Kathleen M. Hagerty and William P. Rogerson (1987), "Robust Trading Mechanisms," *Journal of Economic Theory* 42 (1), 94–107.

[15] Morton I. Kamien, Yair Tauman, and Shmuel Zamir (1990), "On the Value of Information in a Strategic Conflict," *Games and Economic Behavior* 2 (2), 129–153.

[16] David M. Kreps and Robert Wilson (1982), "Sequential Equilibria," *Econometrica* 50 (4), 863–894.

[17] Shengwu Li (2017), "Obviously Strategy-Proof Mechanisms," *American Economic Review* 107 (11), 3257–3287.

[18] Eric Maskin and Jean Tirole (1990), "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values," *Econometrica* 58 (2), 379–409.

[19] Toshihide Matsuo (1989), "On Incentive Compatible, Individually Rational, and Ex Post Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* 49 (1), 189–194.

[20] Paul Milgrom (2011), "Critical Issues in the Practice of Market Design," *Economic Inquiry* 49 (2), 311–320.

[21] Roger B. Myerson and Mark A. Satterthwaite (1983), "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* 29 (2), 265–281.

[22] Parag A. Pathak and Tayfun Sönmez (2008), "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism," *American Economic Review* 98 (4), 1636–1652.

[23] Parag A. Pathak and Tayfun Sönmez (2013), "School Admissions Reform in Chicago and England: Comparing Mechanisms by their Vulnerability to Manipulation," *American Economic Review* 103 (1), 80–106.

[24] Andrés Perea and Jeroen Swinkels (1999), "Selling Information in Extensive Form Games," unpublished paper, Universidad Carlos III de Madrid.

[25] Mark A. Satterthwaite (1975), "Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory* 10 (2), 187–217.

[26] Michael Spence (1973), "Job Market Signaling," *Quarterly Journal of Economics* 87 (3), 355–374.

[27] Takuro Yamashita (2015), "Implementation in Weakly Undominated Strategies: Optimality of Second-Price Auction and Posted-Price Mechanism," *Review of Economic Studies* 82 (3), 1223–1246.