# Informationally Robust Trade and Limits to Contagion

Gabriel Carroll

Department of Economics, Stanford University

579 Serra Mall, Stanford, CA 94305

gdc@stanford.edu

September 1, 2016

**Abstract**

Two agents can each accept or reject a proposed deal, whose value for each agent depends on an unknown state, and may be positive or negative. The deal takes place only if both accept. Each agent can be imperfectly informed, in an arbitrary way, about both her own value and the other agent's. In such environments, contagious adverse selection may prevent the deal from being reached even when it is mutually beneficial ex post. We give an upper bound on the ex-ante expected welfare loss in equilibrium due to such contagion, valid for any information structure. The welfare loss is small if negative values are unlikely ex ante; and under an assumption of known aggregate gains from the deal, our bound is sharp. The bound has a succinct description, even though the equilibrium itself, in any given information structure, may be hard to describe explicitly.

JEL classifications: C72, D82, D83

Keywords: adverse selection, contagion, information structure, robustness

# 1  Introduction

Imagine two agents contemplating a proposed agreement. The value of the agreement for each agent depends on an unknown state of the world, and each of them may have some private information about this state. Ex ante, with high probability, both parties stand to benefit from the deal, but each also foresees some nonnegligible probability that the deal will turn out badly for her. Each agent may accept or reject, and the deal takes place only if both accept.

Many kinds of interactions can be described by this simple setup:

- The agents may be a buyer and seller of a good. Since our model assumes the only options are accept or reject — no bargaining over the price — the applications would be ones where the price is fixed in advance.

  For example, the agents may be a manager and a potential employee, deciding whether to begin an employment relationship. The manager has private information about the nature of the work, and the employee has superior knowledge of her productivity, both of which affect both agents' value for the relationship. The salary is fixed by corporate policy, government pay scale, or union contract, and cannot be negotiated.

- The agents may be representatives of two countries, deciding whether to finalize a trade agreement.

- The agents may be two members of a hiring committee, voting on an offer to a job candidate; the candidate is hired only if the vote is unanimous.

In such a situation, the deal may fail even if its total value to the two agents is positive. Such failure can occur for several reasons. First, of course, either agent can reject the deal if she expects its value to her is negative. Second, there may be *adverse selection*: agent 1 may realize that agent 2 accepts the deal only if it is likely to be favorable to agent 2, which could be an indication that it is unfavorable to agent 1. This can motivate agent 1 to reject, even when her expected value based on her own information alone is positive. A long literature in information economics since Akerlof (1970) has emphasized this possible reason for breakdown of trade.

Third, and more subtly, this breakdown can be exacerbated by *contagion*: Once agent 2 realizes that agent 1 may sometimes reject due to adverse selection as above, this can in

turn make 2 reject for other realizations of her private signal, and so forth. Thus, higher-order beliefs can play an important role in breakdown of trade. A recent literature has started to explore the importance of this contagion in economic outcomes (Rubinstein, 1989; Morris and Shin, 2012; Angeletos and La'O, 2013).

It seems, then, that we cannot predict the outcome of the interaction without knowing the details of the information structure: whether one agent is perfectly informed and the other completely uninformed; or each receives a conditionally-independent noisy signal of the state; or perhaps something much more intricate. Unfortunately, information structures (and especially higher-order information) can be complex, and very hard for an outside observer to model accurately. In this paper, we show that we can nonetheless give a bound on the extent of informational contagion, valid across all information structures. In particular, the effect of contagion goes to zero as the measure of states where either agent is hurt by the deal goes to zero. Moreover, under an assumption that the deal is beneficial in aggregate, our bound is sharp, and identifies the information structure that is most harmful to trade. Perhaps surprisingly, in this worst case, contagion is not a factor at all.

A numerical example will help illustrate our results. Suppose that it is commonly known in advance that the proposed deal produces an aggregate net benefit of 2, but the distribution of this benefit is uncertain. Ex ante, there is a 80% chance that each agent's value is 1; but there is also a 10% chance that agent 1 gains 3 from the deal and agent 2 loses 1, and a 10% chance of the reverse payoffs. In brief, the payoffs from the deal are

$$
\begin{array}{ll}
(3, -1) & \text{with probability } 10\%; \\
(1, 1) & \text{with probability } 80\%; \\
(-1, 3) & \text{with probability } 10\%.
\end{array}
$$

(If the deal is rejected, both agents receive 0.)

In this case, our main theorem will say that, no matter what the information structure is, the Bayesian game between the two agents has an equilibrium in which at least 60% of the aggregate gains from the deal are realized (in expectation). Thus, the effects of contagion are limited to 40% of possible realizations.

Moreover, this 40% bound is sharp. And while we will have to wait to show in detail why the bound holds, sharpness is easy to see: Consider an information structure in which both agents receive the same signal — either a "1-favorable" signal (which indicates that state $(3, -1)$ is relatively likely, in which case agent 2 does not want to trade), or a "2-

3

favorable' 'signal (in which case 1 does not want to trade), or a "normal" signal. The joint distribution of the signal with the players' values is as shown in Table 1; here $\epsilon > 0$ is arbitrarily small. In equilibrium, trade can never occur under the 1-favorable or 2-favorable signals, whose combined probability is $40\% - 2\epsilon$. Note that in this (limiting) worst-case scenario, all information is public and so informational contagion actually plays no role.

| Values | Normal | 1-favorable | 2-favorable |
|--------|--------|-------------|-------------|
| $(3, -1)$ | | 10% | |
| $(1, 1)$ | $60\% + 2\epsilon$ | $10\% - \epsilon$ | $10\% - \epsilon$ |
| $(-1, 3)$ | | | 10% |

Table 1: Distribution of values and (public) signals

Our bound on contagion is both obvious and subtle. It is seemingly obvious in the sense that, if the deal fails, it should be either because agent 1 expects a negative payoff and so rejects the deal, or agent 2 does. In the 10–80–10 distribution above, any event over which agent 1's expected value is negative has probability less than 20%, and likewise for agent 2, thus suggesting the 40% upper bound. But the result is subtle because of the interaction of the agents' information: It is *not* the case that each agent simply refuses the deal when her expected value conditional on her signal is negative. Instead, in equilibrium, each should be conditioning on the information content of the other agent's acceptance. In particular, the equilibrium may require mixed strategies, which makes clear that it cannot be described by a simple yes/no threshold rule on conditional expectations. (An example illustrating this, and adhering to the 10–80–10 distribution, appears in Appendix A.) This means that to prove the bound, we need an argument that avoids actually constructing the equilibrium.

There are two main perspectives on interpretation of our results. One is that this work contributes to theoretical understanding of the role of higher-order beliefs. In particular, the recent literature on contagion uses email-game-like information structures (Rubinstein, 1989; Weinstein and Yildiz, 2007) to show how contagion across information sets can potentially lead to widespread coordination failure. Our results answer this concern by showing that, from an ex-ante point of view, such contagion is limited.

The other interpretation is more positive: our bounds can be useful to parties deciding whether to invest in a future opportunity for mutually beneficial trade. For example, imagine a buyer and seller, making plans at date 0, and anticipating that at date 1 they

4

may wish to trade some widget. At that time, each side may end up with some private information regarding the cost of production and the value to the buyer, and may choose to back out of the proposed trade based on this information. At date 0, they need to decide whether to make some investment (such as production capacity, or building a prototype) that is necessary in order to be able to trade at date 1. Our model applies if they can currently foresee the physical circumstances that affect each party's value for the trade, but cannot anticipate exactly what each party will know when date 1 arrives. A lower bound for the gains from trade realized in equilibrium can potentially provide an immediate guarantee that the investment is worthwhile.[1]

A related application might be to a regulator designing a financial market, in which agents might be able to trade some security whose value depends on future events. If the regulator can anticipate how the events will affect the security's value but not the details of what information traders will have, a lower-bound result can potentially provide assurance that the market can produce enough trade to warrant the fixed costs of creating the market, even though not all socially valuable trades can be realized in equilibrium.

As these examples suggest, our results also tie in with a recent literature on robust mechanism design, and can be interpreted as showing that the simple accept/reject mechanism can provide a guarantee on realizable gains from trade that is as good as any more sophisticated contract design. This connection will be developed more precisely in Section 5.

Methodologically, the broader question behind this paper is: In situations of uncertainty, what can we predict about outcomes of economic interactions without knowing the details of the information structure? Work on this question was pioneered by Bergemann and Morris (2013, 2016), taking a similar approach at an abstract level to general static games, and also applying it to games with a quadratic-normal structure; and by Bergemann, Brooks, and Morris (2013, 2016), performing a similar analysis in a monopoly pricing problem and in a first-price auction. The quadratic-normal application was further advanced by Bergemann, Heumann, and Morris (2015a, 2015b). We contribute to exploration of this general question, choosing one of the simplest possible kinds of games,

---

[1]It might seem that the whole accept-reject game in our model is unnecessary here: if the ex ante gains from trade are positive, the parties should simply contract to trade for sure, and not allow anyone to back out. However, the model fits the following variant: The trade will be socially valuable at date 1 only in a subset of states. At date 1, the parties will know whether the trade is valuable, and in addition, may have further private information about each party's value as above. If the states in which trade is valuable are hard to describe ex ante, the parties may prefer to simply write a contract that allows either of them to back out at date 1. Then our model describes what happens conditional on trade being valuable.

and studying the information-free predictions.

However, a basic difference between this paper and the others mentioned is that the BCE framework used in these other papers considers *all* possible equilibria, across all information structures. In the game we study, it is always an equilibrium for both parties to always reject the deal; and moreover, this equilibrium cannot easily be refined away (see Subsection 4.3). So if we allowed all equilibria, we could not make any predictions about the realized gains from trade. Consequently, we instead study what happens in the best equilibrium for each possible information structure. This difference in focus has a couple consequences. First, the interpretation of the results is different; it makes sense to regard the result directly as a positive prediction of behavior only in situations in which coordination on the best equilibrium is a plausible assumption. Second, we cannot use the same technical tools as in the Bergemann-Brooks-Morris work; we need a new method. This is discussed further at the start of Subsection 3.2 where the proofs appear.

Our contribution is also reminiscent of the work of Kajii and Morris (1997) on ex-ante robustness to incomplete information. They consider Nash equilibria of complete-information games, and give conditions under which any nearby incomplete-information game must have a nearby equilibrium. In fact, their argument leads to a sharp quantitative bound on how far the equilibrium can move when one introduces a given amount of incomplete information. In our model, we give a corresponding quantitative bound on how far the equilibrium can be from the benchmark of full acceptance. A difference is that they allow the payoffs in the newly-introduced states to be adversarially chosen (subject to a bound); we take the payoff distribution as given.

# 2   Model

Let's now flesh out the formal model. The two agents are called 1 and 2. Their values for the proposed deal, $v_1$ and $v_2$, are random variables whose joint distribution is described by an exogenously given probability measure $\mu$ on $\mathbb{R}^2$, with compact support. This $\mu$ describes the prior belief, shared by the two agents and by the outside observer who is trying to make predictions.

Each agent can either accept or reject the deal. If both accept, they receive payoffs $(v_1, v_2)$. If either rejects, then both receive payoff 0. We will assume that neither agent is certain ex ante that the deal is beneficial for her: the events $v_1 < 0$ and $v_2 < 0$ both have positive probability under $\mu$.

Both agents may receive information prior to playing the game, via an *information*

*structure* which is unknown to the observer. We restrict to finite information structures, to avoid complications with equilibrium existence. Thus, an information structure consists of two finite sets of signals, $\mathcal{I}_1$ and $\mathcal{I}_2$, and a joint probability measure $\nu$ on $\mathbb{R}^2 \times \mathcal{I}_1 \times \mathcal{I}_2$, such that the marginal of $\nu$ on the $\mathbb{R}^2$ component coincides with $\mu$. The signals received by the two agents will be denoted by $\eta_1 \in \mathcal{I}_1$ and $\eta_2 \in \mathcal{I}_2$.

Any information structure induces a Bayesian game, in which the two agents observe their signals and then decide whether to agree to the deal. Each agent $i$'s possible (mixed) strategies are functions $\sigma_i : \mathcal{I}_i \to [0,1]$, denoting the probability of agreeing after each signal. The expected payoffs from a strategy profile $(\sigma_1, \sigma_2)$ are

$$u_1(\sigma_1, \sigma_2) = \int \sigma_1(\eta_1)\sigma_2(\eta_2)v_1 \, d\nu, \qquad u_2(\sigma_1, \sigma_2) = \int \sigma_1(\eta_1)\sigma_2(\eta_2)v_2 \, d\nu. \qquad (2.1)$$

(Here and subsequently, whenever we write an integral, the domain of integration is the entire probability space unless indicated otherwise.) The strategy profile is a (Bayesian Nash) *equilibrium* if

$$u_1(\sigma_1, \sigma_2) \geq u_1(\sigma_1', \sigma_2) \qquad \text{and} \qquad u_2(\sigma_1, \sigma_2) \geq u_2(\sigma_1, \sigma_2')$$

for any alternative strategies $\sigma_1', \sigma_2'$.

The observer would like to make robust predictions about the best possible equilibrium. This could naturally be taken to mean the equilibrium that realizes the highest sum of the agents' payoffs; but we could also imagine other criteria, e.g. the highest probability of mutual acceptance, or total welfare with some states weighted more heavily than others. We give a general formulation. So, we assume the observer has an objective, represented by some bounded, measurable function of $v_1, v_2$, call it $w(v_1, v_2)$: The observer gets $w(v_1, v_2)$ when both agents accept and 0 otherwise. Thus, the observer's criterion is

$$W(\sigma_1, \sigma_2) = \int \sigma_1(\eta_1)\sigma_2(\eta_2)w(v_1, v_2) \, d\nu.$$

For example, if we define $w(v_1, v_2) = v_1 + v_2$ then this captures the expected aggregate surplus realized in equilibrium; if $w(v_1, v_2) = 1$ then we have the probability of acceptance.

We then say that a value $x$ for the observer's criterion is a *robust prediction* if, for every information structure $(\mathcal{I}_1, \mathcal{I}_2, \nu)$, there exists an equilibrium $(\sigma_1, \sigma_2)$ satisfying $W(\sigma_1, \sigma_2) \geq x$.[2] We would like to give such a robust prediction and, if possible, identify

---

[2]Admittedly, the term "prediction" is an imperfect fit, since it may suggest a point estimate, whereas

the maximum robust prediction.

Our analysis will also lead us naturally to look at situations where there is no private information, i.e. there may be an informative signal but any such signal is observed by both agents. Explicitly, we say the information structure is *public* if $\mathcal{I}_1 = \mathcal{I}_2$ and the measure $\nu$ places probability 1 on the event $\eta_1 = \eta_2$. We say that a value $x$ is a *robust prediction under public information* if, for every public information structure $(\mathcal{I}_1, \mathcal{I}_2, \nu)$, there exists an equilibrium $(\sigma_1, \sigma_2)$ satisfying $W(\sigma_1, \sigma_2) \geq x$.

We have so far made no assumptions on the exogenous value distribution $\mu$. At some points, we will impose one or the other of the following conditions, under which we can show stronger results:

- **Condition A:** With probability 1 under $\mu$, $\max\{v_1, v_2\} \geq 0$.

- **Condition B:** With probability 1 under $\mu$, $\max\{v_1, v_2\} \geq 0$ and also $w(v_1, v_2) \geq 0$.

In the benchmark case where the observer's criterion is the aggregate welfare $w(v_1, v_2) = v_1 + v_2$, the first part of Condition B is redundant: the condition just requires that $v_1 + v_2 \geq 0$ for sure, i.e. there is ex-ante certainty that the deal is beneficial in aggregate (as in the example from the introduction).

# 3  Results

## 3.1  Measures and decompositions

To identify how good or bad an equilibrium outcome is from the observer's point of view, it suffices to describe the event where the deal is rejected: An upper bound on the size of the rejection event (in the best equilibrium) means a lower bound on the observer's criterion, and thus gives a robust prediction.

Our main results describe these possible rejection events. The first main result says that for any information structure, there exists an equilibrium in which the event of rejection is bounded above by the union of two other events, one on which the expected value of $v_1$ is negative and one on which the expected value of $v_2$ is negative. Under Condition A, we can easily strengthen this to ensure that these two bounding events are disjoint. (Note that this result is just a characterization of the overall rejection event: It

in our language, if $x$ is a robust prediction then any lower value is as well. A name like "robustly attainable value" might be more descriptive. But we keep "prediction" for simplicity.

does *not* say that in equilibrium, agent 1 rejects on the $v_1$-negative event, and agent 2 rejects on the $v_2$-negative event.)

The second main result is a sort of converse: for any event that can be expressed as a disjoint union of two negative-value events of this form, there is an information structure that ensures the deal is rejected there. Moreover, one can choose the information structure to be public.

If $\mu$ satisfies Condition B, then these two results are exactly complementary: taken together, they identify the maximum robust prediction.

The formal statements will require some further definitions. The verbal descriptions so far have been in terms of states of nature and rejection events, but in fact it will be more convenient for us to work with (sub-probability) measures on $\mathbb{R}^2$. Talking in terms of states and events would require explicit reference to probability spaces, which would have to be large enough to describe not only the players' signals but also any coin flips involved in mixed strategies, thus creating extra baggage to carry around. We are not interested in the coin flips themselves, but just in how the possible outcomes — acceptance or rejection — carve up the probability mass that is initially represented by $\mu$. This division can be most succinctly expressed via measures.

Specifically, given an information structure, and given (mixed) strategies $(\sigma_1, \sigma_2)$, we can define a measure $\mu_D$ by

$$\mu_D(E) = \int \sigma_1(\eta_1)\sigma_2(\eta_2) \, \mathbb{1}((v_1, v_2) \in E) \, d\nu,$$

for any measurable set $E \subseteq \mathbb{R}^2$. So for any $E$, $\mu_D(E)$ gives the probability that the pair of values $(v_1, v_2)$ is in $E$ *and* the deal is mutually accepted. (Compare to $\mu(E)$ which simply gives the probability that $(v_1, v_2) \in E$.) Likewise, we can define $\mu_{ND}$ by

$$\mu_{ND}(E) = \int (1 - \sigma_1(\eta_1)\sigma_2(\eta_2)) \, \mathbb{1}((v_1, v_2) \in E) \, d\nu.$$

This is the probability that $(v_1, v_2) \in E$ and at least one agent rejects the deal. We call $\mu_D$ and $\mu_{ND}$ the *deal measure* and *no-deal measure* associated to strategies $(\sigma_1, \sigma_2)$, and note that $\mu_D + \mu_{ND} = \mu$. Note also that we can rewrite the observer's criterion as

$$W(\sigma_1, \sigma_2) = \int w(v_1, v_2) \, d\mu_D. \tag{3.1}$$

Now say that a measure $\mu'$ on $\mathbb{R}^2$ has a *negative decomposition bound* if there exist

two (nonnegative) measures $\mu_1, \mu_2$ with the following properties:

(i) $\mu' \leq \mu_1 + \mu_2$ (that is, $\mu'(E) \leq \mu_1(E) + \mu_2(E)$ for every $E$);

(ii) $\mu_1 \leq \mu$ and $\mu_2 \leq \mu$;

(iii) $\int v_1 \, d\mu_1 < 0$ and $\int v_2 \, d\mu_2 < 0$.

Say that $\mu'$ has a *disjoint negative decomposition bound* if (ii) can be replaced by the stronger

(ii') $\mu_1 + \mu_2 \leq \mu$.

Our first main result is then expressed as follows:

**Proposition 3.1.** *(a) Let $(\mathcal{I}_1, \mathcal{I}_2, \nu)$ be any information structure. There exists an equilibrium $(\sigma_1, \sigma_2)$ whose no-deal measure has a negative decomposition bound.*

*(b) If $\mu$ satisfies Condition A, then for any information structure, there exists an equilibrium whose no-deal measure has a disjoint negative decomposition bound.*

The (partial) converse proposition says that, given any measure that has a disjoint negative decomposition bound, we can find an information structure where the deal is necessarily rejected at least that often in equilibrium.

**Proposition 3.2.** *Let $\mu'$ be a measure that has a disjoint negative decomposition bound. Then there exists an information structure such that, in any equilibrium, the no-deal measure $\mu_{ND}$ satisfies $\mu_{ND} \geq \mu'$. Moreover, we can take this information structure to be public.*

Evidently, Proposition 3.1 leads to an upper bound on the size of the no-deal measure. If the ex-ante probability that $v_1 < 0$ or $v_2 < 0$ is small, then the measures $\mu_1$ and $\mu_2$ in the definition of a negative decomposition bound must be small, to satisfy (iii); and so $\mu_{ND}$, to have a negative decomposition bound, must also be small — that is, the deal is mutually accepted with high probability.

Under Condition B, Propositions 3.1 and 3.2 together pin down the maximum robust prediction that the observer can make:

**Corollary 3.3.** *Suppose Condition B is satisfied. Then, for a real number $x$, the following are equivalent:*

*(a) x is a robust prediction;*

*(b) x is a robust prediction under public information;*

*(c) $x \leq \int w(v_1, v_2) \, d\mu - \sup_{\mu'} \int w(v_1, v_2) \, d\mu'$, where the supremum is over all measures $\mu'$ having a disjoint negative decomposition bound.*

**Proof:** (a) $\Rightarrow$ (b): If a prediction of $x$ is valid for any arbitrary information structure, it is valid for any public information structure.

(b) $\Rightarrow$ (c): Suppose that the conclusion (c) fails to hold. Then $x > \int w(v_1, v_2) \, d\mu - \int w(v_1, v_2) \, d\mu'$ for some particular $\mu'$ that has a disjoint negative decomposition bound. So, by Proposition 3.2, there exists a public information structure where, in every equilibrium, the no-deal measure is at least as large as $\mu'$, and therefore the deal measure $\mu_D$ is bounded above by $\mu - \mu'$. So, by (3.1) and the fact that $w \geq 0$ everywhere, in every equilibrium the observer's criterion is $W(\sigma_1, \sigma_2) \leq \int w(v_1, v_2) \, d(\mu - \mu') < x$. Thus, $x$ is not a robust prediction under public information.

(c) $\Rightarrow$ (a): Suppose $x$ satisfies the condition of (c). Take any information structure. Proposition 3.1(b) applies (since Condition A holds), and we get an equilibrium $(\sigma_1, \sigma_2)$ whose no-deal measure $\mu_{ND}$ has a disjoint negative decomposition bound. Then, using (3.1),

$$W(\sigma_1, \sigma_2) = \int w(v_1, v_2) \, d\mu_D = \int w(v_1, v_2) \, d\mu - \int w(v_1, v_2) \, d\mu_{ND} \geq x$$

by the assumption in (c). So $x$ is a robust prediction. $\square$

Ahead, in Subsection 3.3, we will show how to mechanically calculate the supremum in (c). But first, we prove our main results.

## 3.2 Proofs of main results

We first consider how to prove Proposition 3.1 — existence of a "good" equilibrium for any information structure.

One natural proof approach would be to follow the revelation-principle-style argument familiar from the correlated and Bayes correlated equilibrium literature (Aumann, 1987; Myerson, 1986; Bergemann and Morris, 2016): replace the information structure by one in which each player has just two possible signals, "accept" or "reject," corresponding to the action she would have taken under the original signal; and specify that each player's equilibrium action is to follow her signal. Unfortunately, this simplification will not work

in our setting because we are proving a bound on the *best* equilibrium for each information structure, not on all equilibria. Simplifying the information structure in this way may introduce new, spurious equilibria. So proving a statement about the best equilibrium for the simplified information structure will not tell us anything about the original information structure.[3]

Another approach would be to try to explicitly describe the equilibrium, by applying contagion directly: Start from the strategy profile where both agents always accept the deal. Now identify the information sets (i.e. signals) of each agent where her (expected) value for the deal is negative, and change her strategy to have her reject here. This gives us a new strategy profile. Now adverse selection may apply: there may be new signals where an agent's expected value becomes negative once conditioning on the other agent's acceptance. Change her strategy to reject here, again obtaining a new strategy profile. And so forth. If this process terminates at the equilibrium strategy profile, then perhaps it allows us to infer properties of this profile.

This cannot be exactly correct: As mentioned in the introduction (and detailed in Appendix A), the equilibrium may require mixing, so an iterative algorithm like the above will not actually reach it. But it turns out that a variant does work. The algorithm above considers, at each step, the strategy profile that accepts the deal on all signals not yet reached by the contagion, and rejects on all signals that have been reached. Our variant, instead, considers a Nash equilibrium of the "constrained game" where players are required to accept on the signals not yet reached by contagion, but are free to accept or reject on the signals that have been reached. Such an equilibrium always exists, by the Nash existence theorem. At each step of the iteration, contagion spreads to some new signal where a player would prefer to reject, given the opponent's current strategy. When the iterative process terminates, the signals reached by contagion are not an exact description of the equilibrium no-deal event, but they do give an upper bound for it. By keeping mildly careful accounts of how the contagion spreads, we can show that this upper bound meets the conditions required for Proposition 3.1.

As a historical note, the "constrained game" method to show existence of equilibria with specific properties has been used elsewhere (for example Monderer and Samet, 1989); this author does not know of any previous work where it has been used iteratively, as is done here.

---

[3]The simplification actually does work for the worst-case information structure in the proof of Proposition 3.2 below; no new and better equilibria are introduced. But again, this doesn't help prove Proposition 3.1, because we cannot know this information structure is the worst case until we have already proven the proposition.

**Proof of Proposition 3.1:** First we prove part (a). Take an information structure as given.

We successively define sequences of signal sets $\mathcal{J}_1^k \subseteq \mathcal{I}_1, \mathcal{J}_2^k \subseteq \mathcal{I}_2$ and functions $\lambda_1^k, \lambda_2^k : \mathcal{I}_1 \times \mathcal{I}_2 \to [0,1]$, for $k = 0, 1, \ldots$. These sets and functions will be made to satisfy the following conditions:

(a) $\lambda_1^k(\eta_1, \eta_2) = 0$ whenever $\eta_1 \in \mathcal{J}_1^k$;

(b) $\lambda_2^k(\eta_1, \eta_2) = 0$ whenever $\eta_2 \in \mathcal{J}_2^k$;

(c) if $(\eta_1, \eta_2) \notin \mathcal{J}_1^k \times \mathcal{J}_2^k$, then $\lambda_1^k(\eta_1, \eta_2) + \lambda_2^k(\eta_1, \eta_2) \geq 1$;

(d) if $\mathcal{J}_1^k \neq \mathcal{I}_1$, then $\int \lambda_1^k(\eta_1, \eta_2) v_1 \, d\nu < 0$;

(e) if $\mathcal{J}_2^k \neq \mathcal{I}_2$, then $\int \lambda_2^k(\eta_1, \eta_2) v_2 \, d\nu < 0$.

For each player $i$, $\mathcal{J}_i^k$ will be the set of signal realizations that have not yet been reached by contagion, at stage $k$ of the iterative algorithm. $\lambda_1^k$ and $\lambda_2^k$ will be the accounting that we eventually use to construct the negative decomposition bound.

For the base case, we take $\mathcal{J}_1^0 = \mathcal{I}_1$, $\mathcal{J}_2^0 = \mathcal{I}_2$, and $\lambda_1^0, \lambda_2^0$ identically zero. It is clear that (a) and (b) hold, and (c)–(e) are vacuous.

Now suppose these sets and functions have been defined for some $k$. Consider the Bayesian game where each player learns her signal according to $\nu$, and accepts or rejects the deal, with the constraint that each player $i$ must accept whenever $\eta_i \in \mathcal{J}_i^k$. That is, the (mixed) strategy space of $i$ is the set of $\sigma_i : \mathcal{I}_i \to [0,1]$, such that $\sigma_i(\eta_i) = 1$ whenever $\eta_i \in \mathcal{J}_i^k$; and the payoffs are given by (2.1). This constrained game has a Bayesian Nash equilibrium, call it $(\sigma_1, \sigma_2)$ for now.

Suppose that $(\sigma_1, \sigma_2)$ is not an equilibrium of the original, unconstrained game. In this case we will define $\mathcal{J}_1^{k+1}, \mathcal{J}_2^{k+1}, \lambda_1^{k+1}, \lambda_2^{k+1}$. One of the players has a profitable deviation, say player 1 (the argument for player 2 is analogous). In particular, there is at least one signal $\eta_1^*$ on which she benefits from deviating. That is, there is $\sigma_1'$ that agrees with $\sigma_1$ for all signals except $\eta_1^*$, and such that

$$u_1(\sigma_1', \sigma_2) > u_1(\sigma_1, \sigma_2). \tag{3.2}$$

We must have $\eta_1^* \in \mathcal{J}_1^k$, because otherwise the deviation $\sigma_1'$ would be allowed in the constrained game, and (3.2) contradicts the assumption that $(\sigma_1, \sigma_2)$ was an equilibrium

13

of the constrained game. Therefore, $\sigma_1(\eta_1^*) = 1$, and $\sigma_1'(\eta_1^*) < 1$. So (3.2) implies

$$\int_{\eta_1 = \eta_1^*} \sigma_2(\eta_2) v_1 \, d\nu < 0. \tag{3.3}$$

Define $\mathcal{J}_1^{k+1} = \mathcal{J}_1^k \setminus \{\eta_1^*\}$, and define

$$\lambda_1^{k+1}(\eta_1, \eta_2) = \begin{cases} \sigma_2(\eta_2) & \text{if } \eta_1 = \eta_1^*, \\ \lambda_1^k(\eta_1, \eta_2) & \text{otherwise.} \end{cases}$$

Also define $\mathcal{J}_2^{k+1} = \mathcal{J}_2^k$ and $\lambda_2^{k+1} = \lambda_2^k$.

Thus, stated in words: At step $k+1$, contagion has spread to the one new signal $\eta_1^*$; and we update our accounting by recording agent 2's current strategy $\sigma_2$ in the values of $\lambda_1(\eta_1^*, \eta_2)$, keeping all other $\lambda$'s unchanged.

We check that (a)-(e) are satisfied for step $k+1$. It is straightforward to see that (a) for $k+1$ follows from (a) for $k$. For (c), we only need to check the cases where $\eta_1 = \eta_1^*$. There are two possibilities. If $\eta_2 \notin \mathcal{J}_2^k$, then since $\lambda_1^k(\eta_1, \eta_2) = 0$ (by (a) for $k$), we get

$$\begin{aligned}
\lambda_1^{k+1}(\eta_1, \eta_2) + \lambda_2^{k+1}(\eta_1, \eta_2) &\geq \lambda_1^k(\eta_1, \eta_2) + \lambda_2^{k+1}(\eta_1, \eta_2) \\
&= \lambda_1^k(\eta_1, \eta_2) + \lambda_2^k(\eta_1, \eta_2) && \text{(because } \lambda_2^{k+1} = \lambda_2^k) \\
&\geq 1. && \text{(by (c) for } k)
\end{aligned}$$

If on the other hand $\eta_2 \in \mathcal{J}_2^k$, then $\lambda_1^{k+1}(\eta_1, \eta_2) = \sigma_2(\eta_2) = 1$ already. So (c) holds. For (d), we already know $\int \lambda_1^k(\eta_1, \eta_2) v_1 \, d\nu \leq 0$. And

$$\begin{aligned}
\int \lambda_1^{k+1}(\eta_1, \eta_2) v_1 \, d\nu &- \int \lambda_1^k(\eta_1, \eta_2) v_1 \, d\nu \\
&= \int_{\eta_1 = \eta_1^*} (\lambda_1^{k+1}(\eta_1, \eta_2) - \lambda_1^k(\eta_1, \eta_2)) v_1 \, d\nu \\
&= \int_{\eta_1 = \eta_1^*} \sigma_2(\eta_2) v_1 \, d\nu \\
&< 0
\end{aligned}$$

by (3.3). Finally, (b) and (e) hold since $\mathcal{J}_2^{k+1} = \mathcal{J}_2^k$ and $\lambda_2^{k+1} = \lambda_2^k$.

Now, at each step $k$ of this construction, the sets $\mathcal{J}_1^k, \mathcal{J}_2^k$ become weakly smaller, and one of them becomes strictly smaller. By finiteness, the process must stop at some final step $K$. This can only happen when the constrained equilibrium $(\sigma_1, \sigma_2)$ is an equilibrium

14

of the unconstrained game. This will be the equilibrium claimed in the proposition, so we focus now on this $K$ and these strategies. Let $\mu_{ND}$ be the no-deal measure associated to this equilibrium. We must show the existence of measures $\mu_1, \mu_2$ on $\mathbb{R}^2$ satisfying conditions (i)–(iii) for a negative decomposition bound on $\mu_{ND}$.

Suppose for now that $\mathcal{J}_1^K \neq \mathcal{I}_1$ and $\mathcal{J}_2^K \neq \mathcal{I}_2$. Define $\mu_1, \mu_2$ by

$$\mu_i(E) = \int \lambda_i^K(\eta_1, \eta_2) \mathbb{1}\left((v_1, v_2) \in E\right) d\nu$$

for $i = 1, 2$. Note that for any signals $\eta_1, \eta_2$ we have

$$1 - \sigma_1(\eta_1)\sigma_2(\eta_2) \leq \lambda_1^K(\eta_1, \eta_2) + \lambda_2^K(\eta_1, \eta_2). \tag{3.4}$$

This is because either $(\eta_1, \eta_2) \in \mathcal{J}_1^K \times \mathcal{J}_2^K$ and the left side is zero by definition of the constrained game, or else $(\eta_1, \eta_2) \notin \mathcal{J}_1^K \times \mathcal{J}_2^K$ and then (3.4) follows from condition (c). Now integrating (3.4) over any $E \subseteq \mathbb{R}^2$ gives $\mu_{ND}(E) \leq \mu_1(E) + \mu_2(E)$, which is condition (i) for a negative decomposition bound. Condition (ii) follows from the fact that $\lambda_1^K, \lambda_2^K \leq 1$ everywhere. And (iii) holds because (d) and (e) from the iterative process ensure that

$$\int v_i \, d\mu_i = \int \lambda_i^K(\eta_1, \eta_2) v_i \, d\nu < 0$$

for each $i$.

If $\mathcal{J}_1^K$ is all of $\mathcal{I}_1$ or $\mathcal{J}_2^K$ is all of $\mathcal{I}_2$, then the only difference is that condition (iii) becomes an equality. For example, if $\mathcal{J}_1^K = \mathcal{I}_1$, then $\mu_1$ is the zero measure, so we have $\int v_1 \, d\mu_1 = 0$, instead of $< 0$ which is what we need. In this case, redefine $\mu_1$ by

$$\mu_1(E) = \mu(E \cap \{(v_1, v_2) \mid v_1 < 0\}).$$

By assumption, the event $v_1 < 0$ has positive probability under $\mu$, so $\mu_1$ is a nonzero measure and $\int v_1 \, d\mu_1 < 0$; and conditions (i), (ii) still hold. Similarly, if $\mathcal{J}_2^K = \mathcal{I}_2$ then redefine $\mu_2$ analogously. The new $\mu_1, \mu_2$ now satisfy conditions (i)–(iii).

This proves part (a) of Proposition 3.1. To prove part (b), we show how to transform any negative decomposition bound into a disjoint one, under Condition A.

Let $\mu_1, \mu_2$ be as given by part (a), and suppose that Condition A holds. Let $\mu^\Delta$ denote the positive part of the signed measure $\mu_1 + \mu_2 - \mu$. That is, taking $S^+, S^-$ to be the positive and negative events of $\mu_1 + \mu_2 - \mu$ (given by the Hahn decomposition, see

15

Billingsley (2012, p. 447)), we define

$$\mu^\Delta(E) = \mu_1(E \cap S^+) + \mu_2(E \cap S^+) - \mu(E \cap S^+).$$

Note that $\mu_2 \le \mu$ implies $\mu^\Delta(E) \le \mu_1(E \cap S^+) \le \mu_1(E)$ for each $E$, and likewise $\mu^\Delta(E) \le \mu_2(E)$. Define

$$\widetilde{\mu}_1(E) = \mu_1(E) - \mu^\Delta(E \cap \{(v_1, v_2) \mid v_1 \ge 0\})$$

and

$$\widetilde{\mu}_2(E) = \mu_2(E) - \mu^\Delta(E \cap \{(v_1, v_2) \mid v_1 < 0\}).$$

In particular, $0 \le \mu^\Delta \le \mu_1, \mu_2$ implies that $\widetilde{\mu}_1, \widetilde{\mu}_2$ are nonnegative measures; and $\widetilde{\mu}_1 + \widetilde{\mu}_2 = \mu_1 + \mu_2 - \mu^\Delta$.

We will show that $(\widetilde{\mu}_1, \widetilde{\mu}_2)$ is a disjoint negative decomposition bound for the no-deal measure $\mu_{ND}$. Since $\mu$ and $\mu_1 + \mu_2$ are both upper bounds for $\mu_{ND}$, we have for any $E$:

$$
\begin{aligned}
\mu_{ND}(E) &= \mu_{ND}(E \cap S^+) + \mu_{ND}(E \cap S^-) \\
&\le \mu(E \cap S^+) + \mu_1(E \cap S^-) + \mu_2(E \cap S^-) \\
&= \mu_1(E) + \mu_2(E) - \mu^\Delta(E) \\
&= \widetilde{\mu}_1(E) + \widetilde{\mu}_2(E),
\end{aligned}
$$

showing that condition (i) for a disjoint negative decomposition bound is met. Condition (ii') is met because $\widetilde{\mu}_1 + \widetilde{\mu}_2 = \mu_1 + \mu_2 - \mu^\Delta \le \mu_1 + \mu_2 - (\mu_1 + \mu_2 - \mu) = \mu$. And as for (iii), notice that $\widetilde{\mu}_1$ is obtained from $\mu_1$ by removing mass where $v_1 \ge 0$, and $\widetilde{\mu}_2$ is obtained from $\mu_2$ by removing mass where $v_2 \ge 0$ (since $v_1 < 0$ implies $v_2 \ge 0$ by Condition A). Therefore, (iii) for $(\widetilde{\mu}_1, \widetilde{\mu}_2)$ follows from (iii) for $(\mu_1, \mu_2)$. $\qquad\square$

It now remains to prove Proposition 3.2, on existence of an information structure forcing some amount of rejection. This proof is a very simple construction: Let $(\mu_1, \mu_2)$ be the disjoint negative decomposition bound for the given measure $\mu'$. Then, simply have both players observe whether they end up in $\mu_1$, $\mu_2$, or neither.

**Proof of Proposition 3.2:**  Let $(\mu_1, \mu_2)$ be the disjoint negative decomposition bound for $\mu'$, and put $\mu_0 = \mu - (\mu_1 + \mu_2)$. Let $\mathcal{I}_1 = \mathcal{I}_2 = \{0, 1, 2\}$ be the set of signals. Define the measure $\nu$ on $\mathbb{R}^2 \times \mathcal{I}_1 \times \mathcal{I}_2$ as follows: for any $E \subseteq \mathbb{R}^2$,

$$\nu(E \times \{(0,0)\}) = \mu_0(E), \qquad \nu(E \times \{(1,1)\}) = \mu_1(E), \qquad \nu(E \times \{(2,2)\}) = \mu_2(E),$$

16

and $\nu$ puts zero mass on all other signal pairs. Evidently, $\nu$ is a probability measure whose marginal on $\mathbb{R}^2$ is $\mu_0 + \mu_1 + \mu_2 = \mu$. Thus, we have a public information structure. For any equilibrium $(\sigma_1, \sigma_2)$, we must have $\sigma_1(1)\sigma_2(1) = 0$: If $\sigma_2(1) > 0$, then agent 1 will strictly prefer to reject when she receives signal 1, since her payoff from accepting is $\int \sigma_2(1)v_1 \, d\mu_1 < 0$. Likewise, $\sigma_1(2)\sigma_2(2) = 0$. It follows that the resulting no-deal measure satisfies $\mu_{ND} \geq \mu_1 + \mu_2 \geq \mu'$. $\qquad\square$

## 3.3 Computing the robust prediction

When the distribution of values satisfies Condition B, our Corollary 3.3 characterizes the maximum robust prediction, that is, a tight lower bound on the value of the best equilibrium. In this subsection we show how to actually calculate this tight bound. Evidently, this task is equivalent to calculating the supremum in (c) of the corollary: the maximum total value, as measured by the observer's criterion $w$, that can be packed into a measure with a disjoint negative decomposition bound. Equivalently, this is the maximum total value of $w$ that can be packed into two measures carved out of the prior $\mu$, one on which the expected value of $v_1$ is negative and one on which the expected value of $v_2$ is negative. (Because of the strict inequalities, the worst case is not actually attained; we ignore this for the informal description here.)

It is intuitive that this maximum is found by a greedy algorithm: To form the first measure $\mu_1$, start with all the mass from $\mu$ where $v_1$ is negative; then successively add in remaining mass from $\mu$, proceeding from regions where the ratio $w/v_1$ is highest to where it is lowest, stopping once the average value of $v_1$ so far reaches zero. Likewise with $v_2$ to form measure $\mu_2$.

In the benchmark situation where $w(v_1, v_2) = v_1 + v_2$, the result of this greedy algorithm can be re-expressed as follows: the worst-case $\mu_1$ consists simply of the mass from $\mu$ lying above some upward-sloping line through the origin in the $(v_1, v_2)$-plane; similarly, $\mu_2$ consists of the mass lying below some such line. These regions are shown in Figure 1, where the gray heat map represents the density of the prior distribution $\mu$. (Note we ignore the lower-left half-plane $v_1 + v_2 < 0$ since Condition B implies that values in that half-plane never occur.) The worst-case $\mu_1$ consists of all the mass from $\mu$ lying in the horizontally-hatched region, and $\mu_2$ consists of all the mass lying in the diagonally-hatched region. The slopes of the boundary lines are pinned down by the condition that the integral of $v_1$ over the first region (respectively, $v_2$ over the second) should equal zero. The maximum robust prediction is then given by the integral of $v_1 + v_2$ over the middle,

non-hatched region. If the two hatched regions overlapped, then the maximum robust prediction would be zero — i.e. it may be that the deal is never accepted.
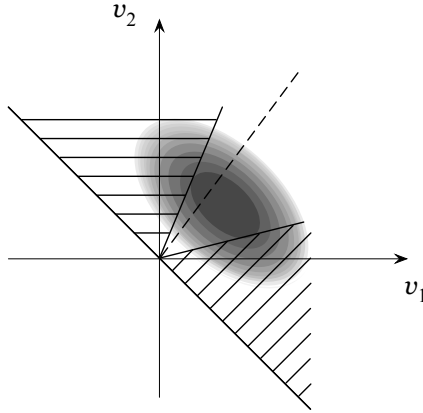


Figure 1: Worst-case measures $\mu_1, \mu_2$ (criterion = aggregate surplus)

For more general criteria $w$, greedily maximizing $w/v_1$ and $w/v_2$ as above is not exactly right because it may lead to measures $\mu_1, \mu_2$ that overlap. In fact, the correct worst-case $\mu_1$ and $\mu_2$ will always be separated in terms of the ratio of the two values: that is, there is some positively-sloped line through the origin, such that $\mu_1$ has support only above the line, and $\mu_2$ has support only below it. (This line need not be unique; an example is shown dashed in Figure 1.) Then, within these respective regions, $\mu_1$ will consist of all the mass from $\mu$ where $v_1$ is negative, plus remaining mass for which the ratio $w/v_1$ is as high as possible; similarly for $\mu_2$, taking mass for which $w/v_2$ is as high as possible.

The rest of this subsection will fill in details; it can be skipped on a casual reading. We first show that the worst-case measures $\mu_1$ and $\mu_2$ can be separated by some positive-slope line as above. Moreover, if there is a positive probability mass lying exactly on the line (a detail omitted from the above description), then this mass can be assigned partially to the region above the line and partially to the region below it, with all points on the line being split in the same proportions.

To write this separation statement explicitly, given $\alpha, \beta \in [0, 1]$, we define subsets of $\mathbb{R}^2$, corresponding to the regions above the dashed line, on the line, and below the line:

$$
\begin{aligned}
E_<^\alpha &= \{(v_1, v_2) \mid v_1 < \alpha(v_1 + v_2)\}, \\
E_=^\alpha &= \{(v_1, v_2) \mid v_1 = \alpha(v_1 + v_2)\}, \\
E_>^\alpha &= \{(v_1, v_2) \mid v_1 > \alpha(v_1 + v_2)\},
\end{aligned}
$$

and define two measures $\overline{\mu}_1^{\alpha,\beta}, \overline{\mu}_2^{\alpha,\beta}$ by

$$\overline{\mu}_1^{\alpha,\beta}(E) = \mu(E \cap E_<^\alpha) + \beta\mu(E \cap E_=^\alpha), \qquad \overline{\mu}_2^{\alpha,\beta}(E) = \mu(E \cap E_>^\alpha) + (1-\beta)\mu(E \cap E_=^\alpha) \quad (3.5)$$

for any $E \subseteq \mathbb{R}^2$. Note that $\overline{\mu}_1^{\alpha,\beta} + \overline{\mu}_2^{\alpha,\beta} = \mu$. (We may omit the superscript $\alpha, \beta$ for brevity.)

We will then say that a pair of measures $(\mu_1, \mu_2)$ is $(\alpha, \beta)$-*separated* if $\mu_1 \leq \overline{\mu}_1^{\alpha,\beta}$ and $\mu_2 \leq \overline{\mu}_2^{\alpha,\beta}$.

**Lemma 3.4.** *Let $(\mu_1, \mu_2)$ satisfy conditions (ii') and (iii) for a disjoint negative decomposition bound. Then there exists a pair $(\widehat{\mu}_1, \widehat{\mu}_2)$ that is $(\alpha, \beta)$-separated for some $\alpha, \beta$, and that also satisfies (ii') and (iii), with $\widehat{\mu}_1 + \widehat{\mu}_2 = \mu_1 + \mu_2$.*

The proof is mechanical: With $(\mu_1, \mu_2)$ given, any choice of parameters $\alpha, \beta$ specifies a way of redividing the mass $\mu' = \mu_1 + \mu_2$ into $\widehat{\mu}_1$ and $\widehat{\mu}_2$. There is some range of $(\alpha, \beta)$ for which the needed inequality $\int v_1 \, d\widehat{\mu}_1 < 0$ is satisfied, and a corresponding range for $\widehat{\mu}_2$; we just need to show that these two parameter ranges overlap. The details are in Appendix B.

Lemma 3.4 shows that in our search for the supremum of $\int w(v_1, v_2) \, d\mu'$ over measures that have a disjoint negative decomposition bound, we can restrict ourselves to decompositions that are $(\alpha, \beta)$-separated for some $\alpha, \beta$.

So, for any given $\alpha, \beta$, define $Y(\alpha, \beta)$ to be the supremum of $\int w(v_1, v_2) \, d(\mu_1 + \mu_2)$ over pairs $(\mu_1, \mu_2)$ that satisfy conditions (ii')–(iii) and are $(\alpha, \beta)$-separated. We just need a way to compute $Y(\alpha, \beta)$ for given $\alpha$ and $\beta$, and then in a subsequent round we optimize over $\alpha, \beta$.

It is evident that

$$Y(\alpha, \beta) = \sup_{\mu_1} \int w(v_1, v_2) \, d\mu_1 + \sup_{\mu_2} \int w(v_1, v_2) \, d\mu_2,$$

where the first supremum is over all measures $\mu_1 \leq \overline{\mu}_1^{\alpha,\beta}$ satisfying $\int v_1 \, d\mu_1 < 0$, and the second is over all measures $\mu_2 \leq \overline{\mu}_2^{\alpha,\beta}$ satisfying $\int v_2 \, d\mu_2 < 0$. We denote these two separate suprema by $Y_1(\alpha, \beta), Y_2(\alpha, \beta)$.

These separate suprema $Y_i$ can each be calculated by the greedy algorithm that takes mass for which $w/v_i$ is as large as possible, up until the point where the total integral of $v_i$ is zero.

Let us give a precise statement. We describe how to compute $Y_1$; then $Y_2$ is analogous.

For $\gamma > 0$ and $\delta \in [0, 1]$, define the subsets of $\mathbb{R}^2$

$$F_<^\gamma = \{(v_1, v_2) \mid v_1 < \gamma w(v_1, v_2)\}, \qquad F_=^\gamma = \{(v_1, v_2) \mid v_1 = \gamma w(v_1, v_2)\}.$$

Then $F_<^\gamma$ is increasing in $\gamma$, and so $\int_{F_<^\gamma} v_1 \, d\overline{\mu}_1$ is also (weakly) increasing in $\gamma$, since the pairs that are in $F_<^\gamma$ but not in $F_<^{\gamma'}$ for $\gamma' < \gamma$ must satisfy $v_1 \geq 0$. The integral is also left-continuous in $\gamma$. Let $\gamma^* \in (0, \infty]$ be the supremum of values such that $\int_{F_<^\gamma} v_1 \, d\overline{\mu}_1 < 0$. (This integral is negative for small enough $\gamma > 0$, so we are assured that $\gamma^* > 0$.) If $\gamma^* < \infty$ then the expression $\int_{F_<^{\gamma^*}} v_1 \, d\overline{\mu}_1 + \delta \int_{F_=^{\gamma^*}} v_1 \, d\overline{\mu}_1$ is weakly increasing in $\delta \in [0, 1]$, and is nonnegative at $\delta = 1$; let $\delta^*$ be the supremum of values for which it is $< 0$. The expression must be equal to 0 at $\delta = \delta^*$.

**Lemma 3.5.** *If $\gamma^* = \infty$ then $Y_1(\alpha, \beta) = \int_{\mathbb{R}^2} w(v_1, v_2) \, d\overline{\mu}_1$. Otherwise,*

$$Y_1(\alpha, \beta) = \int_{F_<^{\gamma^*}} w(v_1, v_2) \, d\overline{\mu}_1 + \delta^* \int_{F_=^{\gamma^*}} w(v_1, v_2) \, d\overline{\mu}_1.$$

*$Y_2(\alpha, \beta)$ can be evaluated by an analogous procedure, using $\overline{\mu}_2$ in place of $\overline{\mu}_1$, and with the roles of $v_1$ and $v_2$ switched in defining $\gamma^*$ and $\delta^*$.*

The proof is in Appendix B.

Finally, we can summarize our work as follows:

**Proposition 3.6.** *Given $\mu$ satisfying Condition B, the following procedure determines the maximum robust prediction:*

1. *For each choice of $\alpha, \beta \in [0, 1]$, split $\mu$ into $\overline{\mu}_1$ and $\overline{\mu}_2$ by (3.5).*

2. *Use the greedy algorithm on this $\overline{\mu}_1$ — taking the mass with the highest ratio $w/v_1$, and all mass where $v_1 < 0$ — to compute $Y_1(\alpha, \beta)$, and likewise with $\overline{\mu}_2$ to compute $Y_2(\alpha, \beta)$, as described in Lemma 3.5. This determines $Y(\alpha, \beta) = Y_1(\alpha, \beta) + Y_2(\alpha, \beta)$ for the given $\alpha$ and $\beta$.*

3. *Then, the maximum robust prediction equals $\int w(v_1, v_2) \, d\mu - \sup_{\alpha, \beta} Y(\alpha, \beta)$.*

**Proof:** The correctness of step 2 follows from Lemma 3.5. The correctness of step 3 then follows from Corollary 3.3 (and the discussion following Lemma 3.4). □

In Section 4, we will give examples to illustrate this procedure.

We note that the brief description given earlier for the aggregate welfare criterion $w(v_1, v_2) = v_1 + v_2$ — as illustrated in Figure 1 — immediately follows as a special case.

## 3.4 Discussion

As we have seen, Proposition 3.1 gives a bound on the no-deal measure, that is tight under Condition B. What happens in general?

If $v_1, v_2$ may simultaneously be negative, the bound implied by the proposition (part (a)) can fail to be tight. To see this, consider any distribution where $v_1 = v_2$ with probability 1. Then the two agents are playing a common interest game, and whatever strategy profile maximizes their expected payoff is an equilibrium. In particular, their payoff in this equilibrium must be at least as high as their payoff from simply always accepting the deal. For an example, suppose that $\mu$ is the distribution consisting of two mass points: $(v_1, v_2) = (-1, -1)$ with probability $2/5$, or $(1, 1)$ with probability $3/5$. Then, this common-interest argument shows that any information structure has an equilibrium where each player's expected payoff is at least $1/5$. But part (a) of the proposition is not strong enough to imply this: in this example, *any* $\mu' \leq \mu$ has a negative decomposition bound, since we can take $\mu_1 = \mu_2$ to be the measure with mass $2/5$ on $(-1, -1)$ and $3/10$ on $(1, 1)$. So part (a) cannot give any nontrivial robust prediction.

The problem is that, in forming a negative decomposition bound, some portion of $\mu$ may be "double-counted" in both $\mu_1$ and $\mu_2$. The proof of part (b) shows that we can get rid of this double-counting as long as $v_1$ and $v_2$ are not both negative; but if they are, we may run into trouble. It seems that the contagion-accounting argument we have used to prove Proposition 3.1 is not strong enough to always give a tight bound.

One more comment: we have also restricted the distribution $\mu$ by assuming there is positive probability both that $v_1 < 0$ and that $v_2 < 0$. But this is a more innocuous assumption, since without it, the game is easy to analyze. For example, if $v_1 \geq 0$ for sure, then we can simply assume that agent 1 always accepts the deal, and 2 best replies; then our tight bound on the no-deal measure is given by the largest $\mu_2$ with $\int v_2 \, \mu_2 < 0$.

# 4 Examples

Here we give a couple of examples illustrating how to compute the maximum robust prediction, following the method in Subsection 3.3. Both of these examples satisfy Condition B, so that the method is valid. We also build further on one of the examples in order to illustrate the resilience of the always-reject equilibrium.

## 4.1 An introductory example

The first application is a simple (perhaps too simple) example adapted from Morris and Shin (2012). It extends the numerical example in the introduction.

The two parties are a buyer and seller, with the opportunity to trade an asset at a fixed price of $p$. It is common knowledge that the asset is worth $v + c$ to the buyer and $v - c$ to the seller, where $v$ is the unknown fundamental value, and $c > 0$ is a known constant. Most likely, $v$ is equal to $p$. However, there is a small probability $\delta$ that the asset is a lemon, with fundamental value $v = p - M$, and probability $\delta$ that it is a peach, with fundamental value $v = p + M$. Here $M$ is a constant with $M > c$.

(In the interpretation given by Morris and Shin, the asset is a mortgage-backed security, and $p$ is a price determined by the market for other similar securities. Any one such security is idiosyncratic and hence there is a possibility of private information. The $\pm c$ trading motive comes from liquidity needs of the buyer and seller. Other interpretations are possible.)

The parties' net gains from the trade are then $(v_1, v_2) = ((v + c) - p, p - (v - c))$. Accordingly, the distribution $\mu$ is:

$$
(v_1, v_2) = \begin{cases} (-M + c, M + c) & \text{(lemon)} & \text{with probability } \delta, \\ (c, c) & \text{(normal)} & \text{with probability } 1 - 2\delta, \\ (M + c, -M + c) & \text{(peach)} & \text{with probability } \delta. \end{cases}
$$

We take $w(v_1, v_2) = 1$ everywhere, so we are interested in robust prediction of the probability of trade. (Taking $w(v_1, v_2) = v_1 + v_2 = 2c$ would give gains from trade.) Note that Condition B is satisfied.

Since the criterion $w(v_1, v_2) = 1$ and the gains-from-trade criterion $w(v_1, v_2) = v_1 + v_2$ are equivalent in this example, the technique illustrated in Figure 1 applies: we form each party $i$'s negative-value measure $\mu_i$ by carving out probability mass from $\mu$ where $i$'s share of the total surplus is as low as possible, up until the point where the expected value of $v_i$ is zero. If these two measures $\mu_1, \mu_2$ end up overlapping then the best robust prediction is no trade.

Specifically, there are two cases depending on parameters:

- If $\delta M/c \leq 1/2$, then the maximal possible total mass of $\mu_1$ is $\delta M/c$ — consisting of the $\delta$ probability of lemon realizations, together with a $\delta(M-c)/c$ probability mass of normal realizations. Likewise the maximal $\mu_2$ consists of the $\delta$ probability of peach and $\delta(M - c)/c$ mass of normal. (Again, these are really suprema, not maxima,

but we glide over this distinction.) Therefore, by Corollary 3.3, the maximum robust prediction is $1 - 2\delta M/c$. That is, for any information structure, there is an equilibrium where the proposed trade occurs with probability at least $1 - 2\delta M/c$; and this bound is tight, even if information is actually public.

To be fully explicit, we describe an information structure approaching the bound: Both parties receive the same signal, $\eta_1 = \eta_2 = \eta \in \{0, 1, 2\}$. The joint distribution of values and signals is as shown in Table 2. Here $\epsilon > 0$ is arbitrarily small.

Under signal 1 — which is a noisy signal of the lemon state — trade cannot occur in equilibrium because player 1 (the buyer) has a negative expected value; and under the peach signal 2, trade cannot occur because player 2 (the seller) has a negative expected value. So trade occurs with probability at most $1 - 2\delta M/c + 2\epsilon$.

| Values $(v_1, v_2)$ | $\eta = 0$ | $\eta = 1$ | $\eta = 2$ |
|---|---|---|---|
| $(-M + c, M + c)$ | $0$ | $\delta$ | $0$ |
| $(c, c)$ | $1 - 2\delta\frac{M}{c} + 2\epsilon$ | $\delta\frac{M-c}{c} - \epsilon$ | $\delta\frac{M-c}{c} - \epsilon$ |
| $(M + c, -M + c)$ | $0$ | $0$ | $\delta$ |

Table 2: Distribution of values and (public) signals

- If $\delta M/c > 1/2$, then the best possible robust prediction is 0: the information may be structured so that no trade can occur in equilibrium.

One possible information structure that yields no trade (not the only one) is to have a publicsignal $\eta \in \{1, 2\}$, jointly distributed with the values as shown in Table 3. Under signal 1, the expected value of $v_1$ is negative; under signal 2, the expected value of $v_2$ is negative.

| Values $(v_1, v_2)$ | $\eta = 1$ | $\eta = 2$ |
|---|---|---|
| $(-M + c, M + c)$ | $\delta$ | $0$ |
| $(c, c)$ | $\frac{1}{2} - \delta$ | $\frac{1}{2} - \delta$ |
| $(M + c, -M + c)$ | $0$ | $\delta$ |

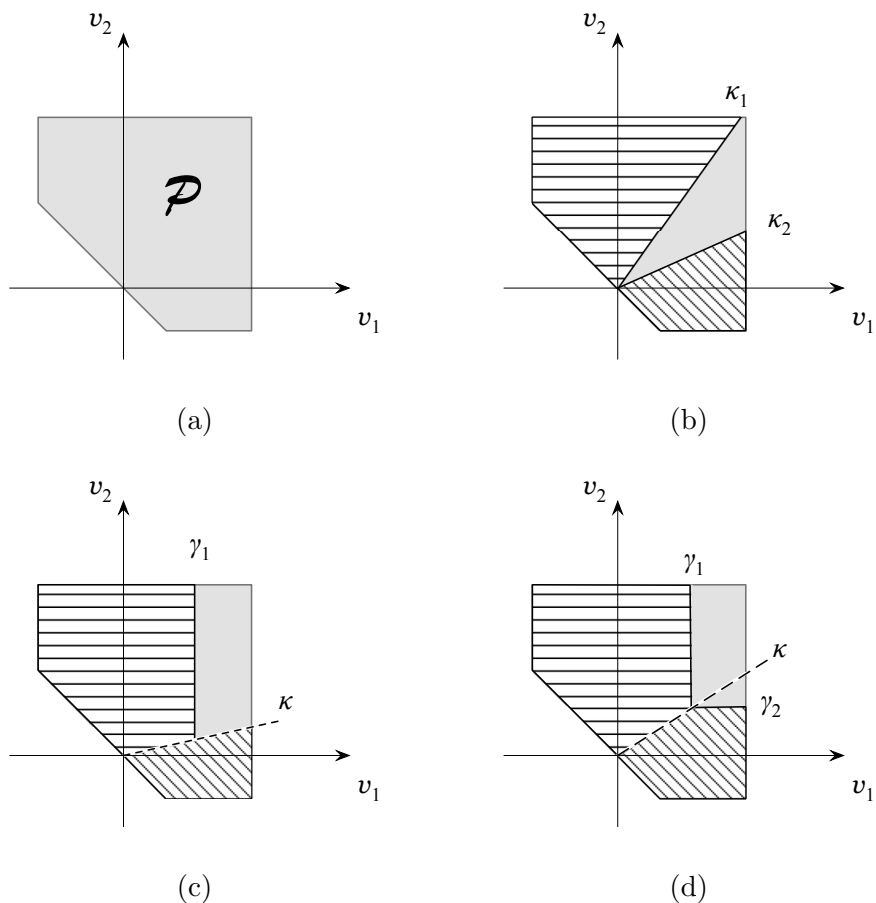Table 3: Distribution of values and (public) signals

Figure 2: Worst-case computation, in an example. (a) The observer's $\mu$. (b) Worst case for the aggregate-welfare criterion. (c) Computing $Y(\alpha)$ for the probability-of-acceptance criterion. (d) Worst case for the probability-of-acceptance criterion.

## 4.2 A more complicated illustration

Here we briefly walk through a more involved example of the method, featuring a continuous and asymmetric distribution of values. To avoid a proliferation of notation, we use specific numbers.

Let $\mu$ be the uniform distribution on the pentagon

$$\mathcal{P} = \{(v_1, v_2) \in \mathbb{R}^2 \mid -2 \leq v_1 \leq 3; -1 \leq v_2 \leq 4; v_1 + v_2 \geq 0\}.$$

This pentagon is shown shaded in Figure 2(a). (We will not attempt to give an interpretation to this distribution.)

We first consider the criterion of aggregate welfare $w(v_1, v_2) = v_1 + v_2$. In this case, again following the method described in Subsection 3.3, the worst-case $\mu_1$ consists of all the mass from $\mu$ lying above some upward-sloping line through the origin; this portion of the pentagon is shown horizontally hatched in Figure 2(b). That is, $\mu_1$ consists of mass distributed uniformly on this region, wth the same density as $\mu$. The slope of this line is determined by the condition that the integral of $v_1$ over this region should be zero. (Unlike in the previous example, we need not worry about what happens with mass exactly on the boundary line, since the mass there is zero.) Likewise, the worst-case $\mu_2$ consists of all the mass from $\mu$ in the region below some other upward-sloping line; this region is shown diagonally hatched in Figure 2(b). The boundary is pinned down by the condition that the integral of $v_2$ over this region is zero.

Thus, in the worst case, both players receive a (public) signal telling them whether $(v_1, v_2)$ has landed in the horizontally-hatched, diagonally-hatched, or gray region. Only in the gray region do both players accept the deal.

To identify the slope of the boundary line for $\mu_1$, write $\kappa_1$ for the slope of this line. This line intersects the upper edge of $\mathcal{P}$ at point $\left(\frac{4}{\kappa_1}, 4\right)$. (This is assuming that the line does indeed intersect the upper edge of $\mathcal{P}$, rather than the right edge; we later check that this is the case.) Thus, the range of possible values of $v_1$ in the horizontally-hatched region is $\left[-2, \frac{4}{\kappa_1}\right]$, and for each such $v_1$ the range of corresponding values of $v_2$ is $[\max\{-v_1, \kappa_1 v_1\}, 4]$. So the condition pinning down $\kappa_1$ is

$$\int_{-2}^{\frac{4}{\kappa_1}} \int_{\max\{-v_1, \kappa_1 v_1\}}^{4} v_1 \, dv_2 \, dv_1 = 0.$$

Some computation leads to $\kappa_1 = \sqrt{2}$. From this we check that $4/\kappa_1 = 2\sqrt{2} < 3$, so that the line does indeed intersect the upper edge of $\mathcal{P}$ as depicted.

By a similar argument, the slope $\kappa_2$ of the line bounding $\mu_2$ is given by the condition

$$\int_{-1}^{3\kappa_2} \int_{\max\{-v_2, \frac{v_2}{\kappa_2}\}}^{3} v_2 \, dv_1 \, dv_2 = 0$$

which gives the solution $\kappa_2 = \sqrt{21}/9$. We check that this line intersects the vertical $v_1 = 3$ at the point $(3, \sqrt{21}/3)$, whose $v_2$-coordinate is $< 4$, so the line does indeed intersect the right edge of $\mathcal{P}$ as shown.

This identifies the worst case. The maximum robust prediction is then given by integrating $v_1 + v_2$, multiplied by the density of $\mu$, over the gray region. The integral

comes to approximately 0.772. Thus, we can robustly predict that for any information structure, there is an equilibrium that realizes expected total welfare of at least 0.772, or 29.1% of the "first-best" where the deal always occurs; and this bound cannot be improved.

We now take the same $\mu$, but consider the probability-of-acceptance criterion, $w(v_1, v_2) = 1$. In this case, we proceed as described in Proposition 3.6. For each $\alpha \in [0, 1]$, we separate $\mu$ into the mass below the line $v_1 = \alpha(v_1 + v_2)$ (i.e. the line through the origin of slope $\kappa = (1 - \alpha)/\alpha$) and the mass above this line. ($\beta$ in the proposition describes how to divide up the mass exactly on the line; again, this is irrelevant in the present example since this mass is zero.) Within the portion above the line, we form $\mu_1$ by taking all the mass with sufficiently low values of $v_1/w$ — that is, with $v_1 < \gamma_1$, for some $\gamma_1$. This is shown by the horizontally-hatched region in Figure 2(c). The value of $\gamma_1$ is determined by the constraint that the integral of $v_1$ over this region should equal 0. If the integral of $v_1$ over the entire region above the line is positive, then no choice of $\gamma_1$ attains this equality, and we instead come as close as possible by taking all the mass above the line to form $\mu_1$. Similarly, $\mu_2$ consists of all the mass below the $\kappa$ line with sufficiently low values of $v_2$, as shown by the diagonally-hatched region. The integral of $v_2$ over this region should equal 0 if possible, and if not possible, then $\mu_2$ consists of all the mass below the line — as shown in the figure in this case. Once $\mu_1$ and $\mu_2$ are constructed, $Y(\alpha)$ consists of the sum of the integrals of $w$ with respect to these two measures — that is, the total area of the two hatched regions (times a constant, the density of $\mu$). Finally, maximizing $Y(\alpha)$ over all $\alpha$ gives the worst case.

The thresholds $\gamma_1$ and $\gamma_2$ are functions of $\alpha$ that cannot be conveniently written in closed form; they are solutions to cubic polynomials whose coefficients depend on $\alpha$. However, we can compute them, and maximize $Y(\alpha)$, numerically. The resulting $\alpha$ is approximately 0.614, i.e. the slope $\kappa$ is 0.629; and the corresponding worst-case $\mu_1$, $\mu_2$ consist of the mass shown in the horizontally- and diagonally-hatched regions of Figure 2(d).[4] The maximum robust prediction is then given by integrating $w$, multiplied by the density of $\mu$, over the remaining, gray region. This gives us 0.165. Thus, for any information structure, there exists an equilibrium where the deal is mutually accepted with probability at least 0.165.

---

[4]As depicted, the boundary lines given by $\gamma_1$ and $\gamma_2$ happen to intersect on the line of slope $\kappa$. This concurrency can be shown to hold more generally, but we do not explore this in detail here.

## 4.3 All-rejection equilibrium

As explained in the introduction, in our game it is natural to look for bounds on the best equilibrium for any given information structure, not on all possible equilibria: the latter question is trivial, because there is always an equilibrium in which both agents always reject the proposed deal. One might try to get rid of such an equilibrium by using a standard refinement, such as elimination of weakly dominated strategies, or more generally trembling-hand perfection. Unfortunately, this does not seem to help. Here is a simple illustration.

We return to the example of Subsection 4.1. Suppose the parameters are such that $\delta M/c$ is small, so that for any information structure there is an equilibrium with a high probability of trade. Now consider the following information structure. The signal sets are $\mathcal{I}_1 = \mathcal{I}_2 = \{L, N, P\}$. The letters stand for "lemon, normal, peach," and the first and last signals are perfectly informative while the middle signal is imperfectly informative. Specifically, conditional on the true value pair $(v_1, v_2)$, the two agents receive signals that are independently drawn from the same distribution, which is given by Table 4.

| Values $(v_1, v_2)$ | $Pr(L \mid v_1, v_2)$ | $Pr(N \mid v_1, v_2)$ | $Pr(P \mid v_1, v_2)$ |
|---|---|---|---|
| $(-M + c, M + c)$ | $1/2$ | $1/2$ | $0$ |
| $(c, c)$ | $0$ | $1$ | $0$ |
| $(M + c, -M + c)$ | $0$ | $1/2$ | $1/2$ |

Table 4: Distribution of each player's signal, conditional on values

For some signal realizations, the players have weakly dominant actions: If player 1 (the buyer) receives $L$, she knows the asset is a lemon for sure (her value for the proposed trade is $-M + c < 0$) so, in any trembling-hand perfect equilibrium, she must reject the trade. If she receives $P$, the asset is a peach for sure, so she must accept. Similarly, in any trembling-hand perfect equilibrium, player 2 must accept on seeing $L$ and reject on seeing $P$.

Consider the following strategy profile: agent 1 accepts *only* upon seeing $P$, and 2 accepts only upon seeing $L$. To check that this is a trembling-hand perfect equilibrium, it suffices to check that each agent is playing a strict best reply when she sees signal $N$. So consider agent 1, say, when she has seen signal $N$. From her point of view, any of the three values for the asset is possible. But 2 will have only agreed to the trade if the asset is a lemon. So agent 1 strictly loses out by accepting the trade; that is, rejecting is

a strict best reply. Similarly for agent 2.

In this equilibrium, the trade is mutually accepted only if agent 1 receives signal $P$ and 2 receives $L$; but this can never happen.

It is worth noting also that, if indeed we do restrict attention to trembling-hand perfect equilibria, our main results are not substantially changed. In particular, the results of Proposition 3.1 remain valid if we require the equilibrium to be trembling-hand perfect, as long as we relax (iii) in the definition of a (disjoint) negative decomposition bound to require weak rather than strict inequalities. (A quick proof sketch: apply the argument in the original proof to the "game with $\epsilon$ trembles," in which each player always accepts with probability $\epsilon$, rejects with probability $\epsilon$, and follows her intended strategy with probability $1 - 2\epsilon$. Then consider a limit of the resulting equilibria and the terminal weights $\lambda_i^K$, as $\epsilon \to 0$.)

# 5   Closing Discussion

We close with some additional discussion of interpretation and connections to literature, and hopes for further work.

## 5.1   Robust mechanism design

Our results can naturally be recast so as to connect with the literature on robust mechanism design (Bergemann and Morris, 2005; Chung and Ely, 2007; Frankel, 2014; Garrett, 2014; Carroll, 2015). In this literature, one envisions a planner designing a mechanism by which agents will interact. As usual, a mechanism specifies actions available to the agents, and an allocation as a function of the actions taken. The planner is unsure about some aspect of the economic environment, and does not represent this uncertainty by a Bayesian prior belief; instead she wants a mechanism that is guaranteed to perform well in all possible environments. This goal is formalized as a maxmin criterion.

In our setting, we can imagine a mechanism design problem where the space of possible "allocations" simply consists of doing the deal or not (or probabilistic mixtures in between). The planner wants to implement the deal, but cannot force either agent to agree — there is an interim individual rationality constraint. The game we have described, where either agent simply chooses to accept or reject, is one possible mechanism; but we could alternatively imagine much more involved interactions.

Formally, a *mechanism* for this setting consists of spaces $M_1, M_2$ of actions for each

28

agent, and a function $g : M_1 \times M_2 \rightarrow [0,1]$, specifying the probability with which the deal is implemented, for each action profile. Interim individual rationality means that each agent should have a "veto" action $m_i \in M_i$ that ensures the deal is not implemented. Thus, for our accept/reject mechanism, $M_1 = M_2 = \{A, R\}$, and $g(A, A) = 1$, $g(A, R) = g(R, A) = g(R, R) = 0$ (and $R$ is veto).

We assume the planner knows the prior $\mu$ on values, but does not know the information structure. We allow the planner to choose a mechanism and, for each information structure $(\mathcal{I}_1, \mathcal{I}_2, \nu)$ (denoted $\nu$ for short), to specify a Bayesian equilibrium $(\sigma_1^\nu, \sigma_2^\nu)$ of the game in which each agent $i$ learns her signal $\eta_i$ and then chooses an action in $M_i$. (Thus the planner gets to choose the equilibrium, as is usual in mechanism design.)

The planner's payoff from information structure $\nu$ is then

$$W^\nu = \int g(\sigma_1^\nu(\eta_1), \sigma_2^\nu(\eta_2)) w(v_1, v_2) \, d\nu.$$

(The equilibrium strategies $\sigma_1^\nu, \sigma_2^\nu$ may involve mixing; here we extend $g$ to mixed actions by linearity.) The planner's uncertainty about the information structure is represented by a worst-case objective: she evaluates any possible mechanism by the infimum of $W^\nu$, over all information structures.

If Condition B is satisfied, then our results quickly imply that the simple accept/reject mechanism is actually optimal for the planner. Indeed, the worst-case information structure in the proof of Proposition 3.2 has the property that *no* mechanism can get the deal accepted with positive probability when signal 1 or 2 is realized. Hence no mechanism can give a performance guarantee better than our robust prediction.

## 5.2 Perspectives

As discussed in the introduction, our results can be interpreted from several perspectives. One is as a purely theoretical comment on the recent literature on informational contagion in games, pointing out limits on the importance of contagion. In particular, under Condition B, informational contagion does not "cause" the socially desirable deal to fail in our game, in the sense that the worst-case information structure does not require any contagion. This contrasts with the basic intuition from adverse selection (Akerlof, 1970) that private information is harmful, and with the more recent applied-theory literature such as Morris and Shin (2012) emphasizing the contributing role of higher-order beliefs. (See also Kessler (2001) or Levin (2001) for another counterpoint to this literature.)

For another, more positive perspective, our results can be useful to parties deciding whether to make an investment that enables future interaction, as in the introductory example of a buyer and seller writing an incomplete contract on possible future trade of a specialized widget. The robust mechanism design interpretation also falls under this heading: it suggests that parties who are uncertain about the future information structure may as well adopt the simple accept/reject mechanism, rather than looking for some more sophisticated alternative mechanism. However, we emphasize that any such positive interpretation depends on the assumption that the players can coordinate on the best Bayesian equilibrium — an assumption that may be reasonable in some applications, such as the contracting or mechanism design cases, but not in others.

A different positive perspective is to see our results as part of the literature on design of information structures (Kamenica and Gentzkow, 2011; Rayo and Segal, 2010; Kolotilin, 2014): the worst-case information structure can be seen as a description of how an adversary might commit to release information so as to try to prevent two agents from reaching a deal.

## 5.3 Multiple proposals

The model here has assumed that there is only a single deal that the two parties can agree to, and we have emphasized applications where this assumption is relevant. But in many other situations, one might imagine that multiple alternative deals are available. For example, in international negotiations, there might be many ways to write a treaty. For a buyer and seller trading a good, there might be many possible prices at which to trade; each price would constitute a different "deal." What would change if we considered such settings, with a mechanism that determines not only whether a deal takes place but which deal is chosen?

In general we would expect very different results. For concreteness, consider the case of a buyer and seller trading a good, with quasi-linear preferences; and for our criterion $w$ consider total realized surplus, which does not depend on the price. Suppose the good is known for certain to be worth more to the buyer than to the seller. With a fixed price as in our model here, Condition B is satisfied, so the worst-case information structure involves no private information. If the price could be endogenously determined then this would no longer be true: With only public information, many simple mechanisms could realize all gains from trade; for example, a mechanism where both agents simultaneously name a price and they trade iff they name the same price. With private information, not

all gains would be realized in general. Indeed much stronger results are known: there are examples in the literature of information structures where *no* mechanism can achieve *any* trade in equilibrium, even though there is common knowledge that the buyer's value is higher than the seller's (Samuelson, 1984; Fieseler, Kittsteiner, and Moldovanu, 2003). So the worst information structure cannot be public.

Can we give good bounds on the surplus attainable with other mechanisms? Unfortunately, the contagion bound argument we have used for the accept/reject mechanism seems difficult to generalize. The argument begins from a strategy profile that is an equilibrium in the "reference" states where both values $v_1, v_2$ are positive, and then changes the strategies one signal at a time, and keeps track of the needed changes in order to obtain a bound on the extent of contagion. For other mechanisms, such as a double auction, it is not clear what reference states or what equilibrium profile to use as a starting point.

The multiple-deal setting certainly raises many natural questions: How would one determine the strongest robust prediction for aggregate welfare if (for example) the parties were able to coordinate on the best mechanism for any given information structure? Is it possible to describe the worst-case information structure, or the optimally robust mechanism? These questions seem significantly harder than the analysis given here for the accept/reject mechanism. We leave them for future work.

# A    An example with mixed strategies

As promised in the introduction, we give here an example showing how the socially optimal equilibrium may involve mixing.

Suppose that the information structure is as follows. Agent 1's signal $\eta_1$ may take one of three possible realizations, which we call $A, B, C$; agent 2's signal $\eta_2$ may take on realizations $X, Y, Z$. The probability of each pair of signals, and the payoffs for each possible signal pair, are as shown in Table 5. The payoffs follow the 10–80–10 example.

The two agents observe their respective signals, then each decide whether to accept or reject the proposed deal.

To identify the (best) equilibrium of this game, the analysis proceeds in three steps:

- If agent 2 receives signal $X$ or $Z$, then for sure she benefits from the deal, so we may as well assume she accepts in these cases.[5] Then, if agent 1 receives signal $A$, then

---

| $\eta_1 \backslash \eta_2$ | $X$ | $Y$ | $Z$ |
|---|---|---|---|
| $A$ | 0.36 | 0.03 | 0.40 |
|  | $1,1$ | $-1,3$ | $1,1$ |
| $B$ | 0.04 | 0.05 |  |
|  | $1,1$ | $-1,3$ |  |
| $C$ |  | 0.10 | 0.02 |
|  |  | $3,-1$ | $-1,3$ |

Table 5: Joint distribution of signals and values

her expected payoff from accepting is positive (although its exact value depends on how agent 2 behaves when she gets signal $Y$). So 1 with signal $A$ accepts as well.

- Does 2 accept the deal when she gets signal $Y$? Suppose she does. Then we can check that agent 1 prefers to reject when she gets signal $B$, and prefers to accept when she gets signal $C$. This fully specifies agent 1's strategy. But then, 2's best reply is to reject when she gets signal $Y$ — a contradiction.

  Now suppose 2 rejects when she gets signal $Y$. Then we can check that 1's best reply is to accept under $B$ and reject under $C$. Given this, 2 prefers to accept under $Y$ — again a contradiction.

- So it must be that in equilibrium agent 2 mixes under signal $Y$. If 2 accepts with probability $q$ under $Y$, then 1's best reply is as follows: under signal $B$, accept if $q < 4/5$ and reject if $q > 4/5$ (and possibly mix at $q = 4/5$); under $C$, reject if $q < 1/15$ and accept if $q > 1/15$ (and mix at $q = 1/15$).

  By examining the possible cases, we soon arrive at the equilibrium: Agent 1 accepts with probability $1/15$ under signal $B$, and 1 under $C$; and agent 2 accepts with probability $4/5$ under $Y$.

In this equilibrium, the ex-ante probability that the proposal is mutually accepted is $667/750 \approx 89\%$ — consistent with the lower bound of $60\%$ given by Proposition 3.1.

---

$Z$. This can only happen in equilibrium if 1 rejects for sure under both $A$ and $B$, or under $A$ and $C$, respectively. Then, the probability that both players accept is at most $60\%$, less than in the equilibrium we compute here. Thus, the equilibrium here is indeed the optimal one.

# B  Omitted Details

**Proof of Lemma 3.4:**  In keeping with the main text, write $\mu' = \mu_1 + \mu_2$. As $\alpha$ ranges over $[0, 1]$, the event $E_<^\alpha$ is increasing in $\alpha$. (This depends on the fact that $v_1 + v_2 \geq 0$ everywhere.) Moreover, any pair $(v_1, v_2)$ contained in one $E_<^\alpha$ but not another satisfies $v_1 \geq 0$, since pairs with $v_1 < 0$ are in every $E_<^\alpha$. Therefore $\int_{E_<^\alpha} v_1 \, d\mu'$ is weakly increasing in $\alpha$. Also, it is negative for small enough $\alpha > 0$, and is left-continuous. Let $\overline{\alpha} \in (0, \infty]$ be the supremum of values for which $\int_{E_<^\alpha} v_1 \, d\mu' < 0$.

Similarly, $\int_{E_>^\alpha} v_2 \, d\mu'$ is weakly decreasing in $\alpha$, negative near $\alpha = 1$, and right-continuous. Let $\underline{\alpha}$ be the infimum of values for which $\int_{E_>^\alpha} v_2 \, d\mu' < 0$.

We show that $\overline{\alpha} \geq \underline{\alpha}$. Suppose not. Then, the sets $E_<^{\overline{\alpha}} \cup E_=^{\overline{\alpha}}$ and $E_=^{\underline{\alpha}} \cup E_>^{\underline{\alpha}}$ are disjoint, except for $(v_1, v_2) = (0, 0)$ which is in both. Define $E_\leq^{\overline{\alpha}} = E_<^{\overline{\alpha}} \cup E_=^{\overline{\alpha}}$ and $E_\geq^{\underline{\alpha}} = (E_=^{\underline{\alpha}} \cup E_>^{\underline{\alpha}}) \setminus \{(0, 0)\}$. These sets are disjoint.

We must have $\int_{E_\leq^{\overline{\alpha}}} v_1 \, d\mu' \geq 0$, otherwise the maximality of $\overline{\alpha}$ would be violated. Similarly, $\int_{E_\geq^{\underline{\alpha}}} v_2 \, d\mu' \geq 0$.

Define two new signed measures by

$$\widetilde{\mu}_1(E) = \mu_1(E) - \mu'(E \cap E_\leq^{\overline{\alpha}}), \qquad \widetilde{\mu}_2(E) = \mu_2(E) - \mu'(E \cap E_\geq^{\underline{\alpha}}).$$

Note that $\widetilde{\mu}_1$ is nonpositive on $E_<^{\overline{\alpha}}$ and nonnegative on $E_>^{\overline{\alpha}}$, hence

$$\int (v_1 - \overline{\alpha}(v_1 + v_2)) \, d\widetilde{\mu}_1 \geq 0.$$

Similarly

$$\int (v_2 - (1 - \underline{\alpha})(v_1 + v_2)) \, d\widetilde{\mu}_2 \geq 0.$$

Then we have

$$0 > \int_{\mathbb{R}^2} v_1 \, d\mu_1 - \int_{E_\leq^{\overline{\alpha}}} v_1 \, d\mu' = \int_{\mathbb{R}^2} v_1 \, d\widetilde{\mu}_1 \geq \overline{\alpha} \int_{\mathbb{R}^2} (v_1 + v_2) \, d\widetilde{\mu}_1,$$

$$0 > \int_{\mathbb{R}^2} v_2 \, d\mu_2 - \int_{E_\geq^{\underline{\alpha}}} v_2 \, d\mu' = \int_{\mathbb{R}^2} v_2 \, d\widetilde{\mu}_2 \geq (1 - \underline{\alpha}) \int_{\mathbb{R}^2} (v_1 + v_2) \, d\widetilde{\mu}_2.$$

So $\int_{\mathbb{R}^2} (v_1 + v_2) \, d\widetilde{\mu}_1 < 0$ and $\int_{\mathbb{R}^2} (v_1 + v_2) \, d\widetilde{\mu}_2 < 0$, and therefore

$$\int_{\mathbb{R}^2} (v_1 + v_2) \, d(\widetilde{\mu}_1 + \widetilde{\mu}_2) < 0.$$

However, $\widetilde{\mu}_1 + \widetilde{\mu}_2$ is a nonnegative measure since

$$(\widetilde{\mu}_1 + \widetilde{\mu}_2)(E) = \mu'(E) - \mu'(E \cap E^{\overline{\alpha}}_{\leq}) - \mu'(E \cap E^{\underline{\alpha}}_{\geq}) = \mu'(E \setminus (E^{\overline{\alpha}}_{\leq} \cup E^{\underline{\alpha}}_{\geq})) \geq 0$$

for any $E \subseteq \mathbb{R}^2$. Since $v_1 + v_2 \geq 0$ $\mu'$-almost everywhere, we have a contradiction.

So indeed we have $\overline{\alpha} \geq \underline{\alpha}$. If $\overline{\alpha} > \underline{\alpha}$, we can take $\alpha$ to be any number in between and $\beta$ to be arbitrary. Then define

$$\widehat{\mu}_1(E) = \mu'(E \cap E^{\alpha}_{<}) + \beta\mu'(E \cap E^{\alpha}_{=}), \tag{B.1}$$
$$\widehat{\mu}_2(E) = \mu'(E \cap E^{\alpha}_{>}) + (1 - \beta)\mu'(E \cap E^{\alpha}_{=}). \tag{B.2}$$

Now

$$E^{\alpha}_{<} \subseteq E^{\alpha}_{<} \cup E^{\alpha}_{=} \subseteq E^{\alpha'}_{<} \cup \{(0,0)\}$$

for any $\alpha' \in (\alpha, \overline{\alpha})$, which readily implies

$$\int_{\mathbb{R}^2} v_1 \, d\widehat{\mu}_1 = \int_{E^{\alpha}_{<}} v_1 \, d\mu' + \beta \int_{E^{\alpha}_{=}} v_1 \, d\mu' < 0,$$

and by a similar argument

$$\int_{\mathbb{R}^2} v_2 \, d\widehat{\mu}_2 < 0.$$

Thus, $(\widehat{\mu}_1, \widehat{\mu}_2)$ satisfies condition (iii) for a disjoint negative decomposition bound. Evidently $\widehat{\mu}_1 + \widehat{\mu}_2 = \mu' = \mu_1 + \mu_2$, so (ii') for the original pair of measures implies (ii') for the new pair. And from the definitions (B.1–B.2) we can see that the new pair is $(\alpha, \beta)$-separated. So we are finished in this case.

We are left with the case $\overline{\alpha} = \underline{\alpha}$. In this case, we fix $\alpha = \overline{\alpha} = \underline{\alpha}$ and repeat the argument with $\beta$.

Since $v_1, v_2 \geq 0$ everywhere on $E^{\alpha}_{=}$, the expression

$$\int_{E^{\alpha}_{<}} v_1 \, d\mu' + \beta \int_{E^{\alpha}_{=}} v_1 \, d\mu' \tag{B.3}$$

is weakly increasing in $\beta \in [0,1]$. Let $\overline{\beta}$ be the supremum of such values for which it is $< 0$. (If it is $\geq 0$ already at $\beta = 0$ then take $\overline{\beta} = 0$.) Note that by continuity in $\beta$, (B.3) is in fact $\geq 0$ at $\overline{\beta}$, except in the corner case where $\overline{\beta} = 1$ and $\alpha = 1$. But in this corner case, the lemma is easily proven. Indeed, we can then take $(\alpha, \beta) = (1, 1)$, and define $\widehat{\mu}_1$ and $\widehat{\mu}_2$ by (B.1–B.2), and the conclusion of the lemma holds: $\int v_1 \, d\widehat{\mu}_1 < 0$ by assumption,

$\int v_2 \, d\widehat{\mu}_2$ must be $< 0$ because $\widehat{\mu}_2$ only places weight on $E^1_>$, where $v_2 < 0$ for sure, and the rest follows as before. Thus, we may assume that the expression (B.3) is $\geq 0$.

Similarly, the expression $\int_{E^\alpha_>} v_2 \, d\mu' + (1 - \beta) \int_{E^\alpha_{\cong}} v_2 \, d\mu'$ is decreasing in $\beta$; let $\underline{\beta}$ be the infimum of values for which it is $< 0$, or $\underline{\beta} = 1$ if no such values exist. The expression is $\geq 0$ there except if $\underline{\beta} = 0$ and $\alpha = 0$, and again this corner case can be disposed of separately.

Now we show that $\overline{\beta} > \underline{\beta}$. Suppose not. Then take any $\beta$ with $\overline{\beta} \leq \beta \leq \underline{\beta}$. Define

$$
\begin{aligned}
\widetilde{\mu}_1(E) &= \mu_1(E) - \mu'(E \cap E^\alpha_<) - \beta \mu'(E \cap E^\alpha_{\cong}), \\
\widetilde{\mu}_2(E) &= \mu_2(E) - \mu'(E \cap E^\alpha_>) - (1 - \beta) \mu'(E \cap E^\alpha_{\cong}).
\end{aligned}
$$

As before, $\widetilde{\mu}_1$ is nonpositive on $E^\alpha_<$ and nonnegative on $E^\alpha_>$, hence

$$
\int (v_1 - \alpha(v_1 + v_2)) \, d\widetilde{\mu}_1 \geq 0,
$$

and similarly

$$
\int (v_2 - (1 - \alpha)(v_1 + v_2)) \, d\widetilde{\mu}_2 \geq 0.
$$

Now

$$
0 > \int_{\mathbb{R}^2} v_1 \, d\mu_1 - \left( \int_{E^\alpha_<} v_1 \, d\mu' + \beta \int_{E^\alpha_{\cong}} v_1 \, d\mu' \right)
$$

(since the first integral is negative by assumption, and the expression in parentheses is just (B.3) at $\beta$, which is $\geq 0$ because we have assumed we are not in the corner case)

$$
= \int_{\mathbb{R}^2} v_1 \, d\widetilde{\mu}_1 \geq \alpha \int_{\mathbb{R}^2} (v_1 + v_2) \, d\widetilde{\mu}_1.
$$

Thus, $\int_{\mathbb{R}^2} (v_1 + v_2) \, d\widetilde{\mu}_1 < 0$. By a similar argument, $\int_{\mathbb{R}^2} (v_1 + v_2) \, d\widetilde{\mu}_2 < 0$. Adding, $\int_{\mathbb{R}^2} (v_1 + v_2) \, d(\widetilde{\mu}_1 + \widetilde{\mu}_2) < 0$. But $\widetilde{\mu}_1 + \widetilde{\mu}_2 = (\mu_1 + \mu_2) - \mu' = 0$ identically — a contradiction.

Thus, $\overline{\beta} > \underline{\beta}$. So we can choose $\beta \in (\underline{\beta}, \overline{\beta})$. Now let $(\widehat{\mu}_1, \widehat{\mu}_2)$ be defined by (B.1–B.2). It is immediate that $\int_{\mathbb{R}^2} v_1 \, d\widehat{\mu}_1$, which is just (B.3), is $< 0$, and similarly $\int_{\mathbb{R}^2} v_2 \, d\widehat{\mu}_2 < 0$. Thus the new pair satisfies condition (iii), and the rest is checked as before. $\qquad \square$

**Proof of Lemma 3.5:** We just prove the formula for $Y_1$.

First suppose $\gamma^* = \infty$. Then $\int_{\mathbb{R}^2} w(v_1, v_2) \, d\overline{\mu}_1$ is clearly an upper bound for $Y_1(\alpha, \beta)$. From the definition of $\gamma^*$, we have $\int_{F^\infty_<} v_1 \, d\overline{\mu}_1 \leq 0$, where $F^\infty_<$ is the event $(w(v_1, v_2) > 0$ or $v_1 < 0)$. If the inequality is strict, we can take $\mu_1 = \overline{\mu}_1|_{F^\infty_<}$ (that is, the measure

defined by $\mu_1(E) = \overline{\mu}_1(E \cap F^\infty_\lessgtr)$ for any $E$). Otherwise, since there is a positive probability of $v_1 < 0$ under $\mu$ (by assumption) and so also under $\overline{\mu}_1|_{F^\infty_\lessgtr}$, then there is also a positive probability of $v_1 > 0$ under $\overline{\mu}_1|_{F^\infty_\lessgtr}$. So we can form $\mu_1$ from $\overline{\mu}_1|_{F^\infty_\lessgtr}$ by removing an arbitrarily small probability mass on such an event. In either case, we obtain $\mu_1$ with $\int_{\mathbb{R}^2} v_1 \, d\mu_1 < 0$ strictly, and $\int_{\mathbb{R}^2} w(v_1, v_2) \, d\mu_1$ arbitrarily close to $\int_{\mathbb{R}^2} w(v_1, v_2) \, d\overline{\mu}_1$.

Now suppose $\gamma^*$ is finite. Define the measure $\widehat{\mu}_1$ by

$$\widehat{\mu}_1(E) = \overline{\mu}_1(E \cap F^{\gamma^*}_<) + \delta^* \overline{\mu}_1(E \cap F^{\gamma^*}_=).$$

So the expression given as the value of $Y(\alpha, \beta)$ in the lemma statement is just $\int w(v_1, v_2) \, d\widehat{\mu}_1$. Also, we know that $\int v_1 \, d\widehat{\mu}_1 = 0$.

We first show that this value is an upper bound on $Y(\alpha, \beta)$. Otherwise, let $\mu_1$ be a measure with higher value of $\int w(v_1, v_2) \, d\mu_1$, still satisfying $\int v_1 \, d\mu_1 < 0$ and $\mu_1 \leq \overline{\mu}_1$. Define a signed measure by $\widetilde{\mu}_1 = \mu_1 - \widehat{\mu}_1$. Then $\widetilde{\mu}_1$ is nonpositive on $F^{\gamma^*}_<$, and nonnegative on $F^{\gamma^*}_>$ (which we define in the obvious way). Therefore,

$$\int_{\mathbb{R}^2} (v_1 - \gamma^* w(v_1, v_2)) \, d\widetilde{\mu}_1 \geq 0.$$

This implies

$$\int_{\mathbb{R}^2} v_1 \, d\mu_1 - \int_{\mathbb{R}^2} v_1 \, d\widehat{\mu}_1 \geq \gamma^* \left( \int_{\mathbb{R}^2} w(v_1, v_2) \, d\mu_1 - \int_{\mathbb{R}^2} w(v_1, v_2) \, d\widehat{\mu}_1 \right).$$

But here the left side is negative, while the right side is positive — a contradiction.

So $\int_{\mathbb{R}^2} w(v_1, v_2) \, d\widehat{\mu}$ is indeed an upper bound on $Y(\alpha, \beta)$. For the reverse direction, note that, as in the $\gamma^* = \infty$ case, the measure $\widehat{\mu}_1$ places some positive probability on the event $v_1 < 0$ (which is contained in $F^{\gamma^*}_<$), and so it must also place positive probability on $v_1 > 0$. By removing an arbitrarily small amount of probability mass with $v_1 > 0$, we get a new measure $\mu_1$ such that $\int_{\mathbb{R}^2} v_1 \, d\mu_1 < 0$ and $\mu_1 \leq \overline{\mu}_1$, and $\int_{\mathbb{R}^2} w(v_1, v_2) \, d\mu_1$ is arbitrarily close to $\int_{\mathbb{R}^2} w(v_1, v_2) \, d\widehat{\mu}_1$. $\qquad \square$

# References

[1] George A. Akerlof (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* 84 (3), 488–500.

[2] George-Marios Angeletos and Jennifer La'O (2013), "Sentiments," *Econometrica* 81 (2), 739–779.

[3] Robert J. Aumann (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica* 55 (1), 1–18.

[4] Dirk Bergemann, Benjamin Brooks, and Stephen Morris (2013), "The Limits of Price Discrimination," *American Economic Review* 105 (3), 921–957.

[5] Dirk Bergemann, Benjamin Brooks, and Stephen Morris (2016), "First Price Auctions with General Information Structures: Implications for Bidding and Revenue," unpublished, Princeton Economic Theory Center Working Paper #071–2015.

[6] Dirk Bergemann, Tibor Heumann, and Stephen Morris (2015a), "Information and Volatility," *Journal of Economic Theory* 158 (B), 427–465.

[7] Dirk Bergemann, Tibor Heumann, and Stephen Morris (2015b), "Information and Market Power," unpublished, Princeton Economic Theory Center Working Paper #070–2015.

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

[9] Dirk Bergemann and Stephen Morris (2013), "Robust Predictions in Games with Incomplete Information," *Econometrica* 81 (4), 1251–1308.

[10] Dirk Bergemann and Stephen Morris (2016), "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics* 11 (2), 487–522.

[11] Patrick Billingsley (2012), *Probability and Measure*, anniversary edition, Hoboken: Wiley & Sons.

[12] Gabriel Carroll (2015), "Robustness and Linear Contracts," *American Economic Review* 105 (2), 536–563.

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

[14] Karsten Fieseler, Thomas Kittsteiner, and Benny Moldovanu (2003), "Partnerships, Lemons, and Efficient Trade," *Journal of Economic Theory* 113 (2), 223–234.

[15] Alexander Frankel (2014), "Aligned Delegation," *American Economic Review* 104 (1), 66–83.

[16] Daniel Garrett (2014), "Robustness of Simple Menus of Contracts in Cost-Based Procurement," *Games and Economic Behavior* 87, 631–641.

[17] Atsushi Kajii and Stephen Morris (1997), "The Robustness of Equilibria to Incomplete Information," *Econometrica* 65 (6), 1283–1309.

[18] Emir Kamenica and Matthew Gentzkow (2011), "Bayesian Persuasion," *American Economic Review* 101 (6), 2590–2615.

[19] Anke S. Kessler (2001), "Revisiting the Lemons Market," *International Economic Review* 42 (1), 25–41.

[20] Anton Kolotilin (2014), "Experimental Design to Persuade," *Games and Economic Behavior* 90, 215–226.

[21] Jonathan Levin (2001), "Information and the Market for Lemons," *RAND Journal of Economics* 32 (4), 657–666.

[22] Dov Monderer and Dov Samet (1989), "Approximating Common Knowledge with Common Beliefs," *Games and Economic Behavior* 1 (2), 170–190.

[23] Stephen Morris and Hyun Song Shin (2012), "Contagious Adverse Selection," *American Economic Journal: Macroeconomics* 4 (1), 1–21.

[24] Roger B. Myerson (1986), "Multistage Games with Communication," *Econometrica* 54 (2), 323–358.

[25] Luis Rayo and Ilya Segal (2010), "Optimal Information Disclosure," *Journal of Political Economy* 118 (5), 949–987.

[26] Ariel Rubinstein (1989), "The Electronic Mail Game: Strategic Behavior under 'Almost Common Knowledge,'" *American Economic Review* 79 (3), 385–391.

[27] William Samuelson (1984), "Bargaining under Asymmetric Information," *Econometrica* 52 (4), 995–1005.

[28] Jonathan Weinstein and Muhamet Yildiz (2007), "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements," *Econometrica* 75 (2), 365–400.