Inference under Selectively Disclosed Data

Ying Gao*

April 14, 2022

Abstract

This paper considers the disclosure problem of a sender who wants to use hard evidence to persuade a receiver towards higher actions. When the receiver hopes to make inferences based on the distribution of the data, the sender has an incentive to drop observations to mimic the distributions observed under better states. We find that, in the limit when datasets are large, it is optimal for senders to play an *imitation strategy*, under which they submit evidence imitating the natural distribution under some desirable target state. The volume of data that the sender can submit must meet a certain standard, a "burden of proof", before the receiver can be persuaded to take a high action. The outcome exhibits partial pooling: senders are honest when either they have little data or the state is good, but they try to deceive the receiver when they have access to a lot of data and the state is bad.

1 Introduction

In order to take appropriate actions, decision-makers rely on data and evidence supplied by self-interested informants, such as companies or individuals. Their informants, however, are often motivated by private concerns about the conclusions the decision-maker draws from the data, and have strong preferences about what action the decision-maker should take. For example, researchers carrying out experiments might aim to support a particular hypothesis, either out of personal bias or because their research is sponsored by an interested party (e.g. soda manufacturers and drug companies). Public companies that release accounting and performance data aim to benefit their shareholders, and therefore generally prefer to disclose data that increases the value of their stock. The amount and specificity of data available to report in both of these cases is increasing as data becomes easier to generate and store – both the models used to analyze experiments, and those used to predict financial outcomes, often take as inputs many individual datapoints over a variety of possible outcomes.

*I thank Drew Fudenberg, Stephen Morris, and Ian Ball for many helpful discussions and comments, and participants at the MIT theory lunch for useful comments and questions.

We model this scenario as a communication game between a sender with known preferences and a receiver. There is a finite number of states, each of which is associated with a distribution over a finite set of outcomes. The sender and the receiver share a common model about the state of the world and state-induced outcome distributions, but the sender observes the dataset of outcomes, while the receiver does not, and has uncertainty both about the state and about how many draws there are. They perform statistical inference given the dataset they are shown, and taking into account the sender's incentives to selectively withhold nonbeneficial datapoints.

The main result is that senders with large amounts of data are dishonest, at the expense of senders with little data, who are not fully believed even though they play honestly. In equilibrium, there is partial pooling, wherein high-data, low-state types of senders imitate the distribution of data generated in better-state, lower-data versions of the world. Descriptively, when datasets are large, the receiver's problem comes down to playing higher actions only after verifying that senders submit a sufficiently large amount of data, conditional on the distribution they target. The key problem is optimally choosing *which* higher-payoff states to target: persuading the receiver that the world may be in one favorable state entails giving away that the world is not in another favorable state. The volume of data necessary to achieve a given payoff – its "burden of proof" – depends on the underlying value of the targeted state, as well as how imitable it is by types in worse states. Relative to full revelation, the resulting partial pooling equilibrium advantages senders with good access to data under bad states at the expense of senders in good states with little data.

The classic unraveling argument of Grossman (1981) and Milgrom (1981) shows that when the sender has known access to data, full revelation obtains. In contrast, our model extends the models of Dye (1985) and Jung and Kwon (1988), in which receivers are unsure if senders hold evidence, so sophisticated receivers cannot fully separate their uncertainty about the whether the sender is informed from their uncertainty about the implications of the data about the state of the world, and partial pooling obtains.

An common assumption in this literature, especially in the case of single-dimensional evidence, is that good outcomes are exogenously specified: it is clear which types of observations are better than others. In real data, often better and worse states of the world lead to observed distributions over the same set of potential outcomes, with the interpretation of each outcome's frequency dependent on the rest of the data. For example, a high price-to-earnings ratio may in and of itself be strong evidence of negative investment outcomes, but its interaction with high capital investment might be stronger evidence of a different state of the world with low present earnings but high future growth. In order to encompass cases like this, we take an arbitrary experiment as a primitive, and consider the incentives for the sender given the endogenous statistical structure. Here, the optimal amount of a particular outcome to disclose varies depending on the availability of observations of other outcomes, but imitation – sending a distribution that exactly mimics the generating distribution of data under a targeted state – never hurts.

The payoff to senders in the limit can be thought of as the outcome of an equilibrium of a game in which datasets are represented by continuous masses. Senders in this game differ in the total mass of their data endowment, but the distribution of data conditional on the state is deterministic. We show the outcome of this continuous-dataset game is equivalent to the limit outcome as N grows large. As a corollary, the value of additional data – that is, an upward shift in the expected amount of data available to senders, fixing the shape of the data-endowment distribution – to the informativeness of communication is vanishing as the expected amount of data goes to infinity, although, in contrast to the case in which all data are disclosed, communication in the limit is not completely informative.

Our predictions hold for the unique equilibrium of the disclosure game that is robust to credible reinterpretations of messaging strategies by coalitions of senders. These equilibria satisfy a "lexicographic optimality" condition: they maximize the payoff to the highest potential-payoff types of senders over the set of all equilibria. The selection criterion is entirely an assumption about the sender's behavior; however, under an additional assumption that the receiver's preferences are single-peaked, it coincides with, and can be thought of as being alternately selected as, the receiver-optimal equilibrium.

The paper is laid out as follows. Section 2 begins by outlining a model of communication with a finite dataset. In Section 3, we solve an example game and introduce the notions of lexicographic optimality and robustness to an inclusive credible announcement. In Section 4, we turn to the continuous-dataset approximation to the communication model, in which we propose and describe the imitation equilibrium outcome. Section 5 gives an algorithm that constructs the lexicographically optimal outcome in the finite-data case, showing that it is also the unique outcome robust to inclusive announcements. Section 6 links games with large datasets to the auxillary continuous-dataset game by showing that the imitation equilibrium outcome in the auxilliary game is the limit of lexicographically optimal equilibrium outcomes when $N \to \infty$. To conclude, Section 7 discusses extensions to continuous state spaces and outcome spaces in the continuum model, as well as substituting exogenous data generation for costly, endogenous data acquisition.

1.1 Review of literature

A number of papers show how the unraveling results of Grossman (1981) and Milgrom (1981) can fail when the sender is endowed with a random amount of evidence. Dye (1985) and Jung and Kwon (1988) model a single-datum case, under which a nonzero probability of senders failing to receive evidence results in pooling between those senders and senders with unfavorable evidence. Subsequent papers, Shin (1994) and Shin (2003), show that sender-optimality of the "sanitization strategy" that reveals only sufficiently favorable evidence extends to games with multiple pieces of evidence, so long as the payoff-relevant state is binary (success vs. failure).

Like us, Dzuida (2011) investigates the question of how senders pool with one another when datasets are large. Dzuida directly assumes a continuum of data, of either positive or negative "arguments" that are evidence of a better and worse states, respectively. Because, in her case, evidence is ordered, her results resemble our findings in a binary-state, binary-observation case, and she focuses on an outcome that coincides with our selected outcome in the large-data limit model. We show that this outcome is, indeed, the limit of outcomes selected by lexicographic optimality in large, but finite, datasets. Both our paper and Dzuida's argue that, relative to the case with a symmetrically informed receiver, the equilibrium is worse for high-state, low-evidence senders, who cannot distinguish themselves from low-state senders. We also extend this insight to more than two states, and to evidence distributed over observations that are not exogenously positive or negative, but whose interpretation depends endogenously both on how well they separate better from worse states. Several papers by e.g. Felgenhauer and Schulte (2014), model the discretionary disclosure of binary evidence with an endogenous and sequential process of data acquisition. We discuss an extension to our model with endogenous information acquisition preceding disclosure, assuming data is acquired at a fixed, exogenous cost.

We select an equilibrium that seems natural because it is robust to senders credibly switching to using messages in a way that benefits all deviators and nobody else. This is related to the idea of neologism proofness and self-signaling sets in Farrell (1993), and, more closely, to the idea of credible announcements in Matthews et al. (1991). Our refinement differs from the notions of announcement-proofness in Matthews et al. in that a "credible *inclusive* announcement" requires that *all* senders who weakly prefer their outcome under a newly announced messaging strategy participate in the announcement, rather than allowing indifferent types to stay out. In our setting, this guarantees the existence of a unique equilibrium that survives announcements, while without imposing this condition there may be none. Hart et al. (2017) and Rappoport (2022) study outcomes of more general evidence games, with the assumption that the receiver's preferences are single-peaked, under which the equilibrium selected by our refinement is the receiver-optimal equilibrium, and, indeed, the outcome of the receiver-optimal mechanism with commitment.[1] Rappoport (2022), in particular, shows that shifts in the type distribution towards types with a greater capacity to imitate others lower the payoff to every type in the receiver optimal equilibrium under single-peaked preferences, by independently establishing that such an equilibrium can be constructed using the same iterative algorithm we use to construct the equilibrium immune to inclusive announcements.

Our results speak to a discussion of persuasion using hard evidence in fields like scientific research and corporate asset management. Shin (2003) applies the sanitation strategy to the disclosure of independent successes in maximizing the market value of corporate stocks; in comparison, we analyze incomplete disclosure strategies when the market uses large datasets to inform more complex models in which inference of the state depends on signals' joint

[1] The receiver-optimal equilibrium outcome is identical to the optimal outcome with commitment in other types of evidence games as well – see Ben-Porath et al. (2019).

distributions. Relatedly, there is a large body of work examining the effects of publication bias that arises due to the systematic omission of negative or inconclusive results. Simonsohn et al. (2014) and Andrews and Kasy (2019) propose methods to identify and correct for the bias induced by selective reporting of scientific findings, using observable distortions in the distribution of reported data (e.g. the "p-curve"). Although these studies do not microfound strategies for data omission, their inference problem is similar to that faced by our receiver.

2 Model

There is a sender (S), who wishes to communicate to a receiver (R) about an unknown state of the world. The receiver is uninformed, and relies on the sender to provide them with evidence in order to make a choice that affects both themselves and the sender. However, the sender's and receiver's incentives are misaligned: the sender's preferred action for the receiver does not depend on the true state, and instead, the sender always wants the receiver to take a higher action (i.e. one that is more beneficial to the sender). Furthermore, the sender is able to drop data as they please: the dataset they submit to the receiver may be incomplete, and the receiver must make inferences assuming that the sender will omit data when it is in their strategic interest.

States and payoffs. The sender and receiver share a common prior $\beta_0(\cdot)$ on the state of the world $\theta \in \Theta \subseteq \mathbb{R}$. The support of the prior — the set of states they consider possible — is finite, $\Theta = \{\theta_1, \ldots, \theta_J\}$, with θ_j increasing in j. I assume that the receiver takes the action $a_r = \mathbb{E}[\theta]$ that matches the expectation of their belief over Θ.[2] In short, the receiver's optimal action is increasing in their expectation of the state of the world.[3]

The sender is wishes the receiver to take as high an action as possible: their payoff is a_r.[4] Their payoff from persuading the receiver to adopt a given belief β is

$$u_s(\beta) = \mathbb{E}_\beta[\theta].$$

Evidence. The private information of the sender comes in the form of hard evidence about the state of the world. In particular, the sender has access to a dataset. Each *datapoint* in the dataset is an observation within a space of outcomes $\mathcal{D} = \{1, \ldots, D\}$, and each state of the world induces a different distribution of observations – when the state is θ_j, the distribution of outcomes of a single experiment is f_j. I assume that, while all f_j share full support over $\mathcal{D}$, they are distinct, so that any two states are distinguishable by the distribution of outcomes they generate.

[2] Note that elements of Θ and actions a_r are assumed to already be appropriately normalized: if the receiver's optimal action is intead $a'_r = h(\mathbb{E}[v(\theta')])$ where v and h are increasing functions, the mappings $\theta = v(\theta')$ and $a_r = h(a'_r)$ renormalize the state and action space to the correct form.

[3] A canonical example of a payoff function that justifies this choice is $u_r = -(a_r - \theta)^2$.

[4] We use this simple form in order to keep notation clear. Indeed, because the receiver never mixes over actions, this can be extended to any monotone function of a_r without any change in the analysis.

The entire dataset consists of a finite collection of i.i.d. draws of f_j. Different senders differ in how much data they can acquire, and ex-ante, the *mass distribution* of data, $g(n)$, is known to both parties, but the true number of observations n is not. Nevertheless, the number of observations possible is assumed to be bounded, and when the support of $g(n)$ is in $\{1,\ldots,N\}$, the sender's dataset, or *type*, is given by

$$t = \frac{1}{N}(n_1,\ldots,n_D),$$

where $t(d) := \frac{n_d}{N}$ is the normalized total mass of experiments in which the outcome is d. The total normalized mass of the dataset is $\frac{n}{N} = \frac{1}{N}\sum_{d=1}^{D} n_d$, and alternately denoted as $|t|$.

We denote the ex-ante probability that the sender will be of type t by $q(t)$, and the posterior over the state conditional on the sender receiving t as $\pi(\cdot|t)$.[5]

Messaging and inference. The receiver does not directly observe the sender's type. Instead, after receiving a dataset, the sender voluntarily submits a message to the receiver, consisting of observations from the dataset. I assume that the sender's access to data determines whether it is feasible to submit a particular body of evidence to the receiver:

Assumption 2.1 *The sender can send any message $m = \frac{1}{N}(\tilde{n}_1,\ldots,\tilde{n}_D)$ that is a* subset *of their dataset ($m \subseteq t$), where*

$$m \subseteq t \;\Leftrightarrow\; m(d) \leq t(d) \;\forall d \in \mathcal{D}.$$

The disclosure game with these parameters is $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^{J}, g)$, with type space $\mathcal{T}_N$ and message space $\mathcal{M}_N$ that are isomorphic to each other, containing all vectors $\frac{1}{N}(n_1,\ldots,n_D)$ with the sum of nonnegative integers $n_1 + \ldots + n_D \leq N$. Irrespective of N, the spaces $\mathcal{T}_N$ and $\mathcal{M}_N$ can be embedded into a *global data space* $\mathcal{F} = [0,1] \times \Delta\mathcal{D}$, consisting of all vectors $(w_1,\ldots,w_D)$ of nonnegative real weights with $\sum_{d=1}^{D} w_d \leq 1$.

In this game, senders can choose which feasible message to send given their type according to a possibly mixed messaging strategy, $\sigma(\cdot|t) : \mathcal{T}_N \to \Delta\mathcal{M}_N$. Their choice of a message is the only means by which they can influence the receiver's action and their own payoffs.

The receiver's belief over states is $\beta \in \Delta\Theta$. After receiving message m, the receiver updates their beliefs according to $\beta(\cdot|m) : \mathcal{M}_N \to \Delta\Theta$. More primitively, though only of indirect consequence to the sender, the receiver holds beliefs, denoted $\beta[\cdot|m]$ with square brackets,

[5] The exact expressions are

$$q(t) = \frac{n!}{\Pi_{d=1}^{D} n_d!} g(n) \sum_{j'} \beta_0(\theta_{j'}) \Pi_{d=1}^{D} f_{j'}(d)^{n_d}, \quad \text{and} \quad \pi(\theta_j|t) = \frac{\beta_0(\theta_j)\Pi_{d=1}^{D} f_j(d)^{n_d}}{\sum_{j'} \beta_0(\theta_{j'})\Pi_{d=1}^{D} f_{j'}(d)^{n_d}}.$$

about the sender's type, which imply their beliefs about the state:

$$\beta(\theta_j|m) = \frac{\sum_{t\in\mathcal{T}_N} \beta[t|m]\pi(\theta_j|t)}{\sum_{t\in\mathcal{T}_N} \beta[t|m]}.$$

PBE and outcomes. Following the convention in signaling games, our base solution concept is perfect Bayesian equilibrium (alternately referred to as "PBE" or "equilibrium").

Importantly, we assume the sender is unable to commit ex-ante to a messaging policy. While playing interim suboptimally under some type realizations in exchange for more lenient inferences for other types may benefit the sender in expectation, there is little incentive for the sender to keep the commitment in either a one-shot setting or when the sender is anonymous in a large population. Thus, we expect the sender under each type to optimize $\sigma(\cdot|t)$ given their anticipation of the receiver's response.

Definition An equilibrium is a pair (σ^*, β^*) where

1. σ^* prescribes the highest-payoff feasible message to a sender of each type:

$$\sigma^*(\cdot|t) \in \arg\max_{m\subseteq t} u_s(\beta^*(\cdot|m)).$$

2. β^* is consistent with Bayesian updating given knowledge that the sender plays to σ^*:

$$\beta^*[t|m] = \frac{q(t)\sigma^*(m|t)}{\sum_{t\in\mathcal{T}_N} q(t)\sigma^*(m|t)} \qquad \text{for all on-path } m,$$

and $\beta^*[t|m] = 0$ if $m \not\subseteq t$.

There is off-path indeterminacy in the receiver's beliefs, so there may be multiple β^*, differing on off-path messages, that jointly form an equilibrium along with a given σ^*. However, we can define

$$\beta_{\sigma^*}[t|m] := \begin{cases} \frac{q(t)\sigma^*(m|t)}{\sum_{t\in\mathcal{T}_N} q(t)\sigma^*(m|t)} & \text{for all on-path } m, \\ \mathbb{1}(\arg\min_{t'\subseteq m} \mathbb{E}_{\pi(\cdot|t')}[\theta]) & \text{for all off-path } m \in \mathcal{F} \end{cases}$$

that, firstly, extends the receiver's inference function to all messages in $\mathcal{F}$, and therefore all of $\mathcal{M}_N$; and secondly, makes all off-path messages minimally attractive for the sender. Given the following lemma, if $(\sigma^*, \beta_{\sigma^*})$ is a PBE, we will often suppress β and call σ^* an equilibrium:

Lemma 2.2 *A strategy σ^* constitutes a PBE with along with some β if and only if $(\sigma^*, \beta_{\sigma^*})$ is a PBE.*

In general, when N is large, the game has many PBE due to self-reinforcing expectations about both off- and on-path play. Rather than using equilibrium as a final solution concept, in the following two sections we will propose a refinement, *lexicographic optimality*, that selects the equilibria we consider most reasonable, due to their robustness to deviations by a coalition of types of senders. We postpone the discussion of details of equilibrium selection until then.

Finally, an *outcome* of an equilibrium is the mapping from a dataset in $\mathcal{F}$ to the payoff[6] that a sender endowed with the dataset receives by best-responding to β_{σ^*},

$$u_{\sigma^*}(t) = \max_{m \in \mathcal{F}: m \subseteq t} u_s(\beta_{\sigma^*}(\cdot|m)).$$

It is straightforward to define the outcome for a type in $\mathcal{T}_N$ to be their equilibrium payoff, and $u_{\sigma^*}(t)$ coincides with this definition for positive-probability types. However, comparing outcomes across games with different type spaces requires comparing payoffs for types that may have zero probability in one or the other type space. We have done so by extending the outcome under σ^* to the global type space $\mathcal{F}$, under the thought experiment: "what payoff would a sender with dataset t obtain if they know they are playing against a receiver who believes they are in equilibrium σ^* of game $\mathcal{G}_N$, even if t is not a possible type in $\mathcal{G}_N$?"

3 Example: 2 states

We start by introducing the main concepts in a simple example with two states, $\Theta = \{\theta_1, \theta_2\}$. The definitions we introduce in the context of this example will apply also to the case when $|\Theta| > 2$.

Suppose that the state θ_2 is good, and the state θ_1 is bad. The receiver's belief is a single number $\beta(\theta_2) \in [0,1]$, and the sender's problem boils down to maximizing $\beta(\theta_2)$. To further simplify the problem, suppose that the domain of f_j is also binary, $\mathcal{D} = \{1,2\}$. Let $f_2(2) = p_2$ and $f_1(2) = p_1$, with $p_2 > p_1$, so that outcome 2 is more likely under state 2 than state 1: in this case, outcome 2 is "better". The space of possible types of the sender, $\mathcal{T}_N$, is illustrated below, with the notation $t = (n_1, n_2)$.

	$(0,N)$					
	$(0,N-1)$	$(1,N-1)$				
n_2	$(0,N-2)$	$(1,N-2)$	$(2,N-2)$			
	$(0,1N-3)$	$(1,N-3)$	$(2,N-3)$	$(3,N-3)$		
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
	$(0,0)$	$(1,0)$	$(2,0)$	$(3,0)$	$\dots$	$(N,0)$
			n_1			

Table 1: $\mathcal{T}_N$ when $\Theta = \{1,2\}$ and $\mathcal{D} = \{1,2\}$.

[6]Since payoffs are monotone in actions, this is equivalent to a mapping from types to the actions induced by their messages in equilibrium.

The set of possible messages, $\mathcal{M}_N$, is identical to the type space. Table 1 above illustrates the set of messages available to type $t = (1, N-2)$ in blue, and the set of types capable of sending message $m = (1, N-2)$ in red.

Whenever $N \geq 2$, there are multiple equilibria. For instance, when $N = 2$, the data mass distribution is $g(0) = g(1) = g(2) = \frac{1}{3}$, the prior is $\beta(2) = \frac{1}{2}$, and the distribution of outcomes is $p_2 = 0.9$, $p_1 = 0.8$, the game has the 3 equilibria in Table 2.

The first equilibrium separates senders into 3 pools, which all obtain different payoffs, while the outcomes of the remaining 2 equilibria are identical, and involve 2 different payoffs, depending on the sender's type. While types $(0,1)$ and $(1,1)$ could obtain a higher payoff than they do in σ_2^* and σ_3^* by separating from the other 3 types, they do not do so, because of adverse beliefs about the receiver's response to message $(0,1)$. In σ_2^*, because $(0,1)$ is off-path, the receiver may believe that the sender's type is $(1,1)$ with high probability if $(0,1)$ is observed, which makes the message unattractive. In σ_3^*, $(0,0)$ is on-path, but played by $(1,1)$ with greater probability than it is played by $(0,1)$, despite the fact that both types are indifferent between playing it and $(0,0)$: this worsens message $(0,1)$ and improves message $(0,0)$, which in turn supports the indifference between the two messages that gives rise to these counterintuitive mixing probabilities. If types $(0,1)$ and $(1,1)$ could together announce to the receiver that they plan to play as in equilibrium σ_1^* and be believed, they would, and would then keep their word, even without commitment.

		(0,2)	(0,1)	(1,1)	(0,0)	(1,0)	(2,0)
σ_1^*	Messages	(0,2)	(0,1)	(0,1)	(0,0)	(0,0)	(0,0)
	Payoffs	1.56	1.49	1.49	1.47	1.47	1.47
σ_2^*	Messages	(0,2)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)
	Payoffs	1.56	1.48	1.48	1.48	1.48	1.48
σ_3^*	Messages	(0,2)	(0,1) and (0,0)	(0,1) and (0,0)	(0,0)	(0,0)	(0,0)
	Payoffs	1.56	1.48	1.48	1.48	1.48	1.48

Table 2: 3 equilibria of $\mathcal{G}_2$ with $p = 0.9$, $q = 0.8$

Definition Given an outcome u_{σ^*}, a set of types T has a *credible inclusive announcement* that they will play a partial strategy $\hat{\sigma}_M$ over message set M for payoff v if

- $\hat{\sigma}_M : M \times T \to \mathbb{R}$ is such that $\sum_{t \in T} \hat{\sigma}_M(m|t) = 1$ for all $m \in M$, $\sum_{m \in M} \hat{\sigma}_M(m|t) = 1$ for all $t \in T$, and $u_s(\beta_{\hat{\sigma}_M}(\cdot|m)) = v$ for all $m \in M$.
- $T = \{t : u_{\sigma^*}(t) \leq v \text{ and } \exists m \in M \text{ s.t. } m \subseteq t\}$, and there is some $t \in T$ with $u_\sigma(t) < v$.

The equilibrium σ_1^* is not vulnerable to such announcements, and has a simple form: senders send as many observations of outcome 2 as they can, and none of outcome 1. Indeed, for all N there exists an immune equilibrium, and in cases where $|\Theta| = |\mathcal{D}| = 2$, it entails disclosing only observations of outcome 2.

Credible inclusive announcements are related to the concept of a credible announcement Matthews et al. (1991), and are distinguished from them in imposing that all types that weakly prefer to obtain v to their equilibrium payoff participate in the announcement if possible, rather than only types that strictly prefer v and some subset of those with a weak preference. We make this distinction because robustness to credible announcements is too strong, in that it often rules out all equilibria: types that are indifferent between a base equilibrium and an announcement may no longer able to obtain their payoff from the base equilibrium once the announcement is made and believed, and such announcements may not correspond to any equilibrium at all. In contrast, a sequence of improvements from credible inclusive announcements can always be used to construct an equilibrium, and indeed, the outcomes of equilibria robust to such announcements are unique, as I show in section 5.[7]

Hart et al. (2017) propose a different refinement for evidence games, *truth-leaning equilibrium*, which, we show in the Appendix, coincides with equilibria robust to credible inclusive announcements whenever the relation "x can imitate y" is a partial order, a condition satisfied by our evidence structure. They show that this is highly related (equivalent in their setting) to receiver-optimal equilibria: the necessary additional assumption is that $u_r(\theta, a_r)$ be single-peaked in a_r. Indeed, in many canonical settings single-peakedness is natural, and the equilibria we study will be both receiver-optimal and robust to sender coordination. However, the latter holds independently of the single-peakedness assumption. We believe that in a setting where the receiver cannot commit, it is natural that the interests of the sender, who has discretion over the data, will determine the equilibrium played, and so we drop the assumption and consider the equilibrium selected from the sender's side.

Finally, we observe that equilibria that are immune to credible inclusive announcements are the same as those with outcomes that satisfy a *lexicographic optimality* condition: across all equilibria, they give the senders who have the highest potential equilibrium payoffs their best possible payoffs, and conditional on this, they also maximize the payoffs to the next-highest-potential-payoff group of senders, and so on.

To state the definition, let $t^+_\sigma(u)$ be the set of possible types that obtain a payoff of at least u under outcome u_σ.

Definition We say $u_\sigma(\cdot)$ **weakly lexicographically dominates** $u_{\sigma'}(\cdot)$ (i.e., $u_\sigma(\cdot) \succeq_l u_{\sigma'}(\cdot)$) if either there exists an element u of

$$U := \{u : t^+_\sigma(u) \setminus t^+_{\sigma'}(u) \text{ is nonempty}\}$$

that is greater than or equal to every element u' of

$$U' := \{u' : t^+_{\sigma'}(u) \setminus t^+_\sigma(u) \text{ is nonempty}\},$$

[7] Bertomeu and Cianciaruso (2018) propose a different refinement for evidence games, based on neologism-proofness, that considers only payoff-improving reinterpretations of one message at a time. In our setting, the highest-value announcement can involve a partial strategy that mixes between multiple messages, and so equilibria survive their refinement that are not robust to credible inclusive announcements.

or U' is empty.

Definition $u_\sigma(\cdot)$ **strictly lexicographically dominates** $u_{\sigma'}(\cdot)$ (i.e., $u_\sigma(\cdot) \succ_l u_{\sigma'}(\cdot)$) if $u_\sigma(\cdot) \neq u_{\sigma'}(\cdot)$ and $u_\sigma(\cdot) \succeq_l u_{\sigma'}(\cdot)$.

Lexicographic dominance defines a partial order on outcomes. When the poset of outcomes has a maximal element, we call it lexicographically optimal:

Definition $u_\sigma(\cdot)$ is **lexicographically optimal** if it strictly lexicographically dominates all other equilibrium outcomes.

Lemma 3.1 *For any finite N, the equilibrium outcome of $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g)$ immune to credible inclusive announcements lexicographically dominates all other equilibrium outcomes.*

The proof of the lemma comes directly from the construction of lexicographically optimal equilibria in Section 5. Lexicographic optimality and lexicographic undominatedness are direct conditions on equilibrium payoffs that can be easier to apply than robustness to credible inclusive announcements. In particular, they will be applicable to outcomes the auxiliary infinite-data model of the next section, in which payoffs to coalitions of types are not in and of themselves well-defined.

We conclude by characterizing the lexicographically optimal outcome for the general case with two states, two outcomes, and arbitrary N. While we stress that there can be multiple strategy profiles achieving the same outcome, the profile given is relatively straightforward: it involves no knife-edge mixing probabilities. The key is that the optimal strategy requires that senders send the best possible amount – which may not be as much as possible – of the high signal only.

Claim 3.2 *When $|\Theta| = |\mathcal{D}| = 2$, one equilibrium that obtains the unique outcome immune to credible inclusive announcements takes the following form:*

- *On-path messages are $\{(0, n_2[k])\}_{k=1}^K$, with $n_2[1] = N$ and*

$$n_2[k] = \arg \max_{n < n_2[k-1]} u_s \left(\frac{\sum_{n_2=n}^{n_2[k-1]} \sum_{n_1=0}^{N-n_2} \pi(\theta_2 | (n_1, n_2)) q((n_1, n_2))}{\sum_{n_2=n}^{n_2[m-1]} \sum_{n_1=0}^{N-n_2} q((n_1, n_2))} \right)$$

  for all $k > 1$.

- *A sender plays the most demanding on-path message they can send:*

$$\sigma^*(t) = (0, \max(n_2[k] : n_2[k] \leq t(2))).$$

4 Modeling infinite data

A natural question is whether, when N grows large, the distribution of outcomes converges. Suppose, for instance, that for a sequence of games with mass distributions $\{g_n\}_{N \in \mathbb{Z}}$, there is a distribution $g_\infty(\mu)$ supported on $\mu \in [0,1]$ such that the likelihood of n observations satisfies $\lim_{N\to\infty} g_N(n) = g(\frac{n}{N})/N$. Then the distribution of types of the sender converges to the distribution in an auxilliary game, in which the conditional distributions of data are still given by $\{f_j\}_{\theta_j\in\Theta}$, but senders' data mass distribution is $g(\mu)$: a sender with a mass μ of datapoints under state θ_j will observe exactly the function μf_j. Although any continuum of data is perfectly informative about the state, the mass of data received by senders of different types may differ, and affect their ability to imitate each other. For instance, if the state is θ_j and they receive a total mass μ of data, then they receive a measure $\mu f_j(1)$ of observations of outcome 1, $\mu f_j(2)$ of observations of outcome 2, and so on. We assume that the density $g(\mu)$ describing the probability of obtaining a measure μ of data is continuous on its support $[0,1]$, and vanishing to 0 at 1. A sender's type is $t = \mu f_j$ when they receive a mass μ of data and the state is θ_j.

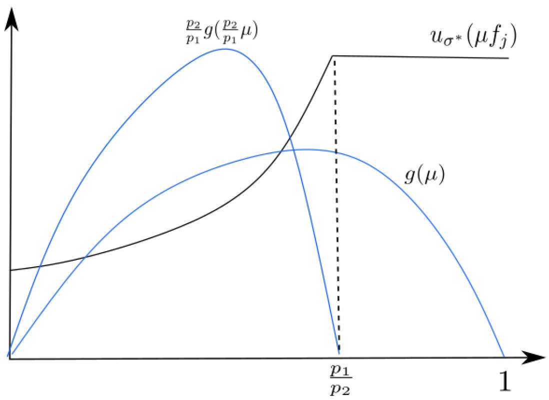
The set of possible types is $\mathcal{T}_\infty = [0,1] \times \Theta \subset \mathcal{F}$, and we let the set of potential messages be $\mathcal{M}_\infty = \mathcal{F}$ – that is, we place no restrictions on what distributions of data the sender may show to the receiver, except that a type t can only send a message m if $m \subseteq t$. Call this infinite-data game $\mathcal{G}_\infty$. Observe that the lexicographic dominance ordering over equilibrium outcomes applies as well to outcomes of equilibria of $\mathcal{G}_\infty$. So, as a first guess to approximating lexicographically undominated outcomes in big-data settings, we may look to equilibria with lexicographically undominated outcomes in $\mathcal{G}_\infty$. In Section 6, we show that indeed, outcomes converge in the limit to the outcome of this auxiliary game.

4.1 A binary-state example

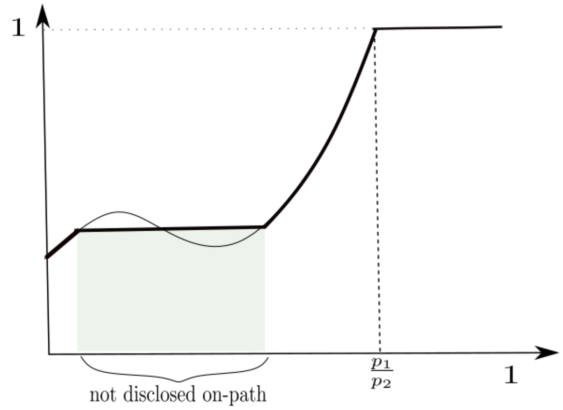
To examine outcomes in the infinite-data approximation, let us return to the example with $|\Theta| = |\mathcal{D}| = 2$. Because $f_1 = (1-p_1, p_1)$ and $f_2 = (1-p_2, p_2)$, types take the form $\mu(1-p_1, p_1)$ or $\mu(1-p_2, p_2)$. Since outcome 2 is better proof that the state is 2 than outcome 1 is, let us focus on equilibria in which, like in the finite-N case, senders disclose only observations of outcome 2.

Figure 1a shows that when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is monotone in μ, disclosing more observations of 2 is always better: conditional on observing a mass μp_2 of observations of 2, the receiver believes the sender either has a mass μ of data and the state is 2, or the sender has a mass $\frac{p_2}{p_1}\mu$ of data and the state is 1, so the sender's payoff is $1 + \dfrac{g(\mu)}{\frac{p_2}{p_1} g(\frac{p_2}{p_1}\mu) + g(\mu)}$.

Equilibrium outcomes must always be monotone: when $t \subseteq t'$, then $u_{\sigma^*}(t) \leq u_{\sigma^*} t'$, since all messages available to t are also available to t'. However, it is possible for $g(\frac{p_2}{p_1}\mu)/g(\mu)$ to be nonmonotone in μ. In this case the strategy "disclose as many observations of 2 as possible" does not respect payoff monotonicity. Instead, the lexicographically undominated outcome involves *ironing* the putative payoff function $1 + \dfrac{g(\mu)}{\frac{p_2}{p_1} g(\frac{p_2}{p_1}\mu) + g(\mu)}$. Figure 1b gives an illustration.



(a) Payoffs as a function of μ and f_j when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is monotone.



(b) Payoffs when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is non-monotone.

Note how the construction coincides with the equilibrium of Claim 3.2: when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is monotone, payoffs to disclosing increasing amounts of outcome 2 are increasing as well. On the other hand, when $\frac{g(\frac{p_2}{p_1}\mu)}{g(\mu)}$ is not, they are not, and in the finite-data equilibrium, the set of on-path messages $\{(0, n_2[k])\}_{k=1}^{K}$ is a strict subset of $\{(0, n)\}_{n=0}^{N}$, with some types pooling with types that have fewer observations of 2.

Finally, note that there is indeterminacy in the equilibrium strategies that would implement u_{σ^*}. Unlike in the finite-data setting, senders could just as well have imitated the entire distribution f_2 by sending μf_2 instead of sending only $(0, \mu p_2)$, since they prove the same thing: the same set of types under both state 1 and state 2 are capable of sending either. We call equilibria in which all on-path messages take the form μf_j *imitation equilibria*.

The imitation equilibrium is special in that, conditional on the mass of outcome 2 disclosed, senders disclose as much of outcome 1 as possible without submitting evidence *against* outcome 2. Intuitively, if μ_2 is the mass of 2 submitted and μ_1 is the mass of 1 submitted, then as long as $\frac{\mu_2}{\mu_1} \geq \frac{p_2}{p_1}$, either a state-2 sender or a state-1 sender with at least a measure μ_2 of outcome 2 can send the dataset, and so μ_1 is uninformative. Otherwise, μ_1 restricts the set of state-2 senders possible, relative to state-1 senders.

It is straightforward to extend the construction of the imitation equilibrium outcome with $|\Theta| = |\mathcal{D}| = 2$ to construct the imitation equilibrium outcome when the state is binary, but the space of experimental outcomes is an arbitrary finite set. Where $\frac{p_2}{p_1}$ gives ratio of the maximum measure of data distributed f_2 that a sender has under state 2 to the measure that a sender endowed with the same total amount of data under state 1 has, the same ratio can be constructed for arbitrary outcome spaces: we define for any particular observation the relative likelihood under distributions f and f' to be

$$LR(f, f'|d) = \frac{f(d)}{f'(d)}$$

and the maximum of $LR(f_j, f_{j'}|d)$ over all d to be

$$r_{j'}(j) = \max_d \left(\frac{f_j(d)}{f_{j'}(d)} \right).$$

Then the equilibrium outcome constructed, replacing $\frac{p_2}{p_1}$ by $r_1(2)$, for the general case is the analogous imitation equilibrium, and is also lexicographically undominated (and, in fact, lexicographically optimal).

4.2 Imitation with $J > 2$ states

We now extend a characterization of an imitation equilibrium outcome to the case with $J > 2$ states. In particular, we look for an equilibrium in which payoffs under every state are a continuous function of μ. While imitation equilibria are not unique, the construction of the continuous-payoff equilibrium follows the intuition that we prioritize awarding high payoffs to senders that have high potential payoffs in equilibrium. Here, we focus on characterizing the equilibrium and its outcome, but we will show in Section 6 that when a sequence of finite models converges to an infinite-data game, the limit of the corresponding lexicographically optimal equilibrium outcomes must converge to exactly the outcome of this imitation equilibrium.

The central object defining the imitation equilibrium is a "burden of proof" associated with each payoff and state, which gives the volume of data imitating the given state distribution that is necessary to obtain the desired payoff. Senders endowed with different datasets will best meet the burden of proof in different ways. Indeed, the equilibrium can be summarized by a vector-valued *burden-of-proof function*, $\hat{\mu}(u) = (\hat{\mu}_1(u), \ldots, \hat{\mu}_J(u))$, such that each sender need only consider the maximal level of utility u such that they can meet the burden of proof for some component j of the associated vector. Their optimal strategy is then to imitate state j using a measure $\hat{\mu}_j(u)$ of data. Correspondingly, the payoff obtained by disclosing (μf_j) is $u_j(\mu)$, which is the (continuous) inverse of $\hat{\mu}$ in that

$$\hat{\mu}_j(u_j(\mu)) = \min\{\mu' : u_j(\mu') = u_j(\mu)\} \qquad \text{and} \qquad u_j(\hat{\mu}_j(u)) = u.$$

where $\hat{\mu}_j(u)$ may also be empty if there is no $\mu \in [\underline{\mu}, \bar{\mu}]$ such that $u_k(\mu) = u$. Indeed, the domain of $\hat{\mu}_j$ will turn out to be $[u_s(\mathbb{1}_1), u_s(\mathbb{1}_j)]$ – imitating the state-j distribution never yields a greater payoff than having the state thought to be j for sure.

As in the case of a binary state, the set set $\{r_j(k)\}_{j,k \in 1,\ldots,J}$ fully characterizes the *pairwise* comparisons between $f_1, \ldots, f_J$, which are the only relevant features for masquerading across states, as they encode how advantaged the data distribution under each state is in imitating another based on their relative similarity. More specifically, all that matters is $\{r_j(k)\}_{j<k}$, as senders under better states will never imitate worse states in equilibrium.

Theorem 4.1 *There exists a unique*[8] *lexicographically undominated equilibrium outcome, and it is implemented by a vector-valued burden of proof function $\hat{\mu}(u) : [0, \theta_J] \to \mathbb{R}^J$ such that*

1. *Its inverse $u_k(\mu)$ is continuous and (weakly) increasing in μ for all k, and $u_k(\max_{j'<j} \frac{1}{r_{j'}(k)}) \geq \theta_j$ for all j and k.*
2. *There is a strategy σ^{im} with $\sigma^{im}(\mu f_j)$ supported on $\{\hat{\mu}_k(u_k(\mu/r_j(k)))f_k : \theta_k \in A_j(\mu)\}$ where*
$$A_j(\mu) = \{\theta_k : k \in \arg\max_k u_k(\frac{\mu}{r_j(k)})\}$$
with $\sigma^{im}(\hat{\mu}_k(u)f_k|\hat{\mu}_k(u)f_k) = 1$ for all k such that $\theta_k \geq u$ and such that for each u and k,
$$u_s\left(\beta_{\sigma^*}(\cdot|\hat{\mu}_k(u)f_k)\right) = u.$$

Then σ^{im} the corresponding equilibrium sender strategy profile.

In the Appendix, we give the full step-by-step construction of σ^{im}. Intuitively, the construction notes that for a target utility level $u \in (\theta_j, \theta_{j+1})$, strategies must consist of senders imitating a state in $j+1, \ldots, J$, so the burden of proof can be projected down to $J - j$ dimensions. The burden of proof is then constructed from the top down, starting with the payoff frontier associated with the highest possible payoff, θ_J. Fixing a preexisting frontier $\hat{\mu}_k(v)$, there are two cases: either $u_k(\hat{\mu}_k(v))$ is strictly increasing in μ_k for all k, or there is a positive-measure set of types that, when pooled, generate a payoff v for the sender, so $u_k(\hat{\mu}_k(v))$ is constant over some range of μ_k. In the former case, the set of types that target some message $\hat{\mu}_k(v)f_k$ are those just barely capable of meeting the kth component of $\hat{\mu}(v)$, but that cannot meet any component of $\hat{\mu}(v)$ with slack: that is, states under which senders are relatively advantaged in sending f_k relative to other components of burden of proof vector $\hat{\mu}(v)$ are the states that target θ_k to obtain a payoff of v. The construction of the continuation of $\hat{\mu}(v)$ in this case comes from the fact that, for any collection of target states such that some senders mix between targeting each of them, the relative rate of change of $u_k(\cdot)$ is pinned by mixing types' indifference. In the latter case, which corresponds to cases in which payoffs from imitating the greatest possible volume of some more favorable state are nonmonotone, senders may send strictly less data distributed f_k than they can. Similarly to the finite-data case, when there exists a positive-measure set of types that attains the highest payoff out of all such type sets, then it is possible to find a set of messages and strategies for senders in the positive-measure type set such that all types included receive the same payoff, which must be v.

Figure 3 shows the result of this process in a setting with 3 states. For any number of states, the resulting sender strategy profile is always part of an equilibrium of the disclosure

[8] Unique up to (outcome-irrelevant) indeterminacy when no amount of data distributed f_j would convince the sender to award a payoff of u.

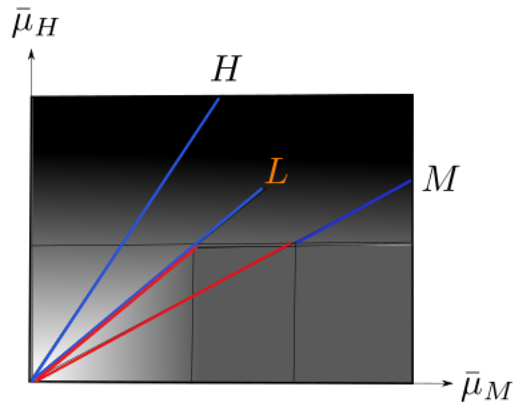


Figure 2: An illustration of sender's disclosure policy in the equilibrium σ^{im} with 3 states: high (H), medium, (M), and low (L). A blue line represents types who masquerade under the high-state distribution; a red line represents types that masquerade under the medium-state distribution; and the coexistence of both denotes mixing.

game. To see why, observe that when the receiver believes off-path messages are negative signals (i.e. are sent by the worst type they could be sent by), then it is necessary to match some dimension of burden-of-proof $\hat{\mu}(u)$ in order to obtain a payoff u. Therefore, the best that any sender in $\tilde{S}_j(u)$ can do is indeed to send a measure $\hat{\mu}_j(u)$ of f_j.

As in the case of binary states, there are many equilibria and not all are likely. In addition to being lexicographically undominated, the one proposed here has the appealing feature that all senders are either truthful, or achieve a higher payoff than they would if their identity was known; this contrasts with equilibria in which senders refrain from sending even positive off-path information, for fear of it being interpreted unfavorably.

A striking feature of the imitation equilibrium outcome is that there are thresholds of data plentifulness, one above which agents lie, and one below which agents, though truthful, are not trusted. The intuition is that, unless the state observed by the sender is actually θ_k, having enough data to send a large volume of distribution f_k implies that the sender also has enough data to obtain a payoff strictly higher than θ_k by targeting a different state. In other words, under every state, if any senders are untruthful, it is those endowed with the most data. Unless the state is the worst possible, a positive-measure set of senders with moderate amounts of data are both "truthful" and are believed: the receiver is sure about the state of the world after observing a message sent by one of these types.[9] The senders with the least data also disclose, as much as possible, data supporting the true state, but the receiver fails to find them credible, and punishes the lack of evidence with skepticism.

[9]Truthfulness, here, means they send a dataset that proves the same thing as if they submitted all their data. We cannot say more than this because of strategic indeterminacy; nevertheless, all such reports signify the same thing in context, and lead to the same outcome.

Theorem 4.2 *Under the equilibrium σ^{im}, there are thresholds $z_j^* > z_j^{**} \geq 0$ for each state such that:*

- *Whenever the sender's type is μf_j with $\mu > z_j^*$, the sender masquerades as a higher type, and receives a payoff $u_{\sigma^{im}}(\mu f_j) > \theta_j$.*
- *Whenever $\mu \in (z_j^{**}, z_j^*]$, the sender is honest and the receiver knows it upon receiving the data: $u_{\sigma^{im}}(\mu f_j) = \theta_j$.*
- *Whenever $\mu \leq z_j^{**}$, the sender is honest, but the receiver believes they are a worse type with positive probability, and $u_{\sigma^{im}}(\mu f_j) < \theta_j$.*

5 Construction of equilibria in $\mathcal{G}_N$

In this section, we give an algorithm to construct equilibria of $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g)$ that are lexicographically optimal and immune to credible inclusive announcements. In doing so, we show that the two refinements select the same, unique, outcome.

We begin with some useful notation. First, fix abstractly a set of types $T \subset \mathcal{T}_N$. Given $\mathcal{T}$, define for every message $m \in \mathcal{M}_N$ and set of messages $M \subseteq \mathcal{M}_N$

$$T^+(m) = \{t \in T : m \subseteq t\} \quad \text{and} \quad T^+(M) = \bigcup_{m \in M} T^+(m),$$

the set of types in T capable of sending $\tilde{f}$ or any $\tilde{f} \in M$, respectively.

Denote the receiver's belief over states after updating their prior based on knowledge that the sender's type is in set T by

$$\beta(\theta|T) = \frac{\sum_{t \in T} \pi(\theta|t) q(t)}{\sum_{t \in T} q(t)}.$$

We say a set of messages $M = \{m_1, \ldots, \tilde{m}_I\}$ implements a pool of sender types $T_{\hat{\sigma}_M}$ if there is an associated partial strategy $\hat{\sigma}_M : M \times \mathcal{T} \to \mathbb{R}$ with $\hat{\sigma}(\cdot|t) \in \Delta M$, satisfying:

A. $t \in T_{\hat{\sigma}_M}(m_i) > 0$ only if $m_i \subseteq t$.

B. $\sum_i \hat{\sigma}_M(m_i|t) = 1$ for all $t \in T_{\hat{\sigma}_M}$.

C. $\sum_{t \in T_{\hat{\sigma}_M}} \hat{\sigma}_M(m_i|t) = 1$ for all $m_i \in M$.

D. $u_s(\beta_{\hat{\sigma}_M}(\cdot|\tilde{f}_i)) = u_s(\beta_{\hat{\sigma}_M}(\cdot|\tilde{f}_j))$ for all i, j.

The payoff to a pool is $u(T_{\hat{\sigma}_M}) := u(\beta(\cdot|T_{\hat{\sigma}_M}))$. Note that types in $T_{\hat{\sigma}_M}$ do not pool in the traditional sense of sending the exact same message (and thus being indistinguishable to the receiver). Instead, they may indeed send different messages that induce different beliefs over the mixture of types; however, these beliefs will result in the receiver taking the same action, and are therefore outcome-equivalent.

Finally, with reference to type set T, define the set of *upper pools* to be the collection of message sets that implement the pooling of the set of all types in T capable of sending them.

$$\mathcal{P}_T = \{M \subseteq \mathcal{M} : M \text{ implements the pooling of } T^+(M)\}$$

Fixing the strategy of the receiver, if we let M be the set of messages such that the receiver's response yields payoff u^* to the sender, and let T be the set of senders incapable of sending any message that yields payoff greater than u^*, then the best response of all senders in $T^+(M)$ to the receiver's strategy is to play some message in M. If, in addition, $M \in \mathcal{P}_T$ and $u(T^+(M)) = u^*$, then there exists a best response by senders in $T^+(M)$ that preserves the payoff to M when the receiver best-responds in turn to the updated strategy.

Lemma 5.1 *For every message set M, there is an upper pool M' consisting of types $T_{\hat{\sigma}_{M'}} \subseteq T^+(M)$ such that $u(T_{\hat{\sigma}_{M'}}) \geq u(T^+(M))$; the inequality is strict if M is not itself an upper pool.*

Lemma 5.2 *For any $T \subseteq \mathcal{T}_N$, the set of utility-maximizing upper pools in $\mathcal{M}_N$, i.e. $\arg\max_{M \in \mathcal{P}_T \bigcup \mathcal{M}_N} u(T_{\hat{\sigma}_M})$, is an upper semilattice in the inclusion order on the set of participating types.*

Observe that it is possible to construct a strategy profile in the following way.

Algorithm (Finite N).

1. Let $T_1 = \mathcal{T}_N$, and define $\mathcal{P}_{T_1}$ to be the set of upper pools over T_1. Find the upper pool in $\mathcal{P}_{T_1} \bigcap \mathcal{M}_N$ that yields the highest payoff to participating senders:
$$M_1 \in \arg \max_{M \in \mathcal{P}_{T_1} \bigcap \mathcal{M}_N} u(T_1^+(M)).$$
If there are multiple such pools, then we take their union, which is also in $\mathcal{P}_{T_1} \bigcup \mathcal{M}_N$ by Lemma 5.2.
2. For $s = 2$ onwards, restrict the set of types to $T_s = T_{s-1} \setminus T_{s-1}^+(M_{s-1})$, and find (the union of)
$$M_s \in \arg \max_{M \in \mathcal{P}_{T_s} \bigcap \mathcal{M}_N} u(T_s^+(M)).$$
3. Continue until $T_s \setminus T_s^+(M_s) = \emptyset$, and define σ^* by $\sigma^*(m|t) = \hat{\sigma}_{M_s}(m)$ where M_s is the pool containing m.

Theorem 5.3 *σ^* is an equilibrium.*

The theorem is immediate from the following lemma, which states that payoffs to the iteratively-constructed pools are strictly decreasing.

Lemma 5.4 *$u(T_m^+(M_m)) > u(T_{m+1}^+(M_{m+1}))$ for all m.*

Indeed, a necessary and sufficient condition for σ^* to be an equilibrium is that $u(T_m^+(M_m)) \geq u(T_{m+1}^+(M_{m+1}))$ for all m; it is additionally true in this case that the inequality is strict, so each successive pool obtains a different payoff.

By construction, u_{σ^*} is unique[10] and lexicographically optimal among equilibrium outcomes. Indeed, all equilibria can be constructed via a version of the algorithm in which the pool chosen in each step need not be the maximal-payoff upper pool. By imposing that we take the largest maximal-payoff pool, we ensure that if $u_{\sigma^*_{alt}}$ coincides with u_{σ^*} for all payoffs greater than v, then the set of types obtaining payoff v is at least as large under u_{σ^*} as it is under $u_{\sigma^*_{alt}}$.

As a proof of Lemma 3.1, we show that u_{σ^*} is the unique outcome of equilibria immune to credible inclusive announcements.

Proof By construction, if $u_{\sigma^*_{alt}} \neq u_{\sigma^*}$, then there exists a v such that the set of pools achieving a payoff greater than v is identical in $u_{\sigma^*_{alt}}$ and u_{σ^*}, but the pool of types T achieving payoff v under u_{σ^*} is a strict superset of that under $u_{\sigma^*_{alt}}$. Then types in T can make a credible inclusive announcement that they will play as they do in σ^*.

If $u_{\sigma^*_{alt}}$ is not lexicographically optimal, then a sequence of improvements by credible inclusive announcements, starting with the senders that achieve the highest payoffs under σ^*, will terminate in an equilibrium with the lexicographically optimal outcome.

A caveat to this method of construction is that the algorithm is highly demanding for large datasets, as its runtime is exponential in N. The equilibrium in the auxiliary infinite-data game is easier to characterize due to a key simplification: the sender's dataset becomes deterministic given the state and the mass of data they receive, and randomness in individual draws ceases to matter. We next turn to a limit result that shows that for large N, outcomes are approximated by the lexicographically undominated outcomes in $\mathcal{G}_\infty$.

6 Convergence of lexicographically optimal equilibria

A sequence of games of finite data, $(\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g_N))_{N=1}^\infty$, converges to $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty)$ if $Ng_N(\lfloor N\mu \rfloor)$ converges uniformly to $g(\mu)$.

Definition A sequence of equilibria $(\sigma_1, \sigma_2, \ldots)$ of games $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g_N))_{N=1}^\infty$ has *outcomes that converge* to the outcome of an equilibrium σ_∞ of the limit infinite-data game $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty)$ if the payoffs $u_{\sigma_N}(t)$ converge uniformly to $u_{\sigma_\infty}(t)$ over $\mathcal{T}_\infty$.

[10] Again, a caveat here is that, while the equilibrium *outcome* that is lexicographically optimal and immune to credible inclusive announcements is unique, in corner cases there can be multiple equilibria that implement it, differing only in the mixing probabilities of different types in the same pool; so we refrain from saying the equilibrium itself is unique.

Note payoffs are only required to converge for types that are possible in the limit, which is consistent with the fact that lexicographic optimality does not constrain payoffs for types that occur with probability (density) 0.

Nevertheless, here they do, and they converge to the imitation equilibrium σ^* of the limit infinite-data game, as our 2nd main theorem shows.

Theorem 6.1 *If σ^{im} is the imitation equilibrium in $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty)$, then along any sequence $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g_N))_{N=1}^\infty$ that converges to $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty)$, the LD equilibrium outcomes converge to $u_{\sigma^{im}}$.*

The full proof is in the Appendix, and a sketch is as follows. When N is very large, there is a type in $\mathcal{T}_N$ close to any type $t \in \mathcal{T}_\infty$. Under the algorithm that generates σ_N^*, that type must obtain the payoff of the maximal upper pool at the step m in which its payoff is assigned. For any ϵ, define

$$T(m) = \{t' \in \mathcal{T}_N : t' \text{ remains at step } m \text{ and } u_{\sigma^{im}}(t') \geq u_{\sigma^{im}}(t) - \epsilon\}.$$

The payoff to the maximal upper pool is lower-bounded by the receiver's belief about the state conditional on the sender being in $T(m)$, since there remains a set of messages M such that all types in $T(m)$ can send at least one message in M, but no types outside $T(m)$ can do so. When the receiver forms their belief about the state conditional on the sender being in $T(m)$, the sender's payoff is bounded below, with the bound approaching $u_{\sigma^{im}}(t) - \epsilon$ as $N \to \infty$; intuitively, this comes from the fact that whenever a set of types in the neighborhood of $\mu f_j \in \mathcal{T}_\infty$ is in the possible set, a corresponding measure of types in the neighborhood of the type that μf_j imitates under the imitation equilibrium must also be in the set. Since the imitated types correspond to better states, the belief given the set of types must be at least as favorable. In the limit as $N \to \infty$, no equilibrium outcome of $\mathcal{G}_N$ can be unilaterally better for all senders in $\mathcal{T}_\infty$, due to Bayes plausibility, and they cannot be worse for any sender, so the two outcomes must coincide.

The convergence result shows that the descriptive features of the lexicographically undominated imitation equilibrium with infinite data – including the targeting strategies wherein types with worse states prove they have sufficient data corresponding to a better state, and including the thresholds for a provided dataset to be fully credible – are also features of lexicographically dominant equilibria in the limit with large, but finite, datasets. Work by Rappoport (2022) shows that, for finite n, an upward shift in the receiver's beliefs about $g(n)$ worsens outcomes for each type of sender, due to the receiver's increased skepticism (although an upward shift of the distribution of evidence may not worsen the sender's *ex-ante* outcomes). Since infinite-data case is the limit as N grows large, and upward shifts in the distribution $g(\mu)$ induce upward shifts in $g(n)$, his results and the convergence result together imply that a similar comparative static holds for both large n and the infinite-data lexicographically undominated outcome: when $g(\mu)$ shifts upwards, for a sender with fixed μ and f_j, payoffs decrease, and the thresholds z_j^* and z_j^{**} of 4.2 are monotone increasing with upwards (i.e. FOSD) shifts of g.

7 Extensions

Continuous state and outcome space in $\mathcal{G}_\infty$

For tractability in the case where N is finite, we have assumed that Θ and $\mathcal{D}$ are discrete and finite. In the auxiliary infinite-data model, these assumptions can be relaxed. In particular, when Θ is finite, equilibrium in $\mathcal{G}_\infty$ is determined only by the finite set of comparisons $\{r_j(k)\}_{j<k}$ regardless of D, and it is possible to consider an infinite, continuous space of possible observations $D \subseteq \mathbb{R}^N$ without any change in the analysis.

In some of the existing literature, most notably Dzuida (2011), Θ is a continuum. Dzuida assumes that (binary) observations satisfy MLRP over the 1-dimensional state space. Our model, in the case of MLRP, nests Dzuida's setup in the case of no honest types, and predictions coincide. If Θ and D are compact subsets of $\mathbb{R}$ with upper and lower bounds $\bar{\theta}, \underline{\theta}$ and $\bar{d}, \underline{d}$ respectively, then if MLRP holds,

$$r_\theta(\theta') = \frac{f_{\theta'}(\bar{d})}{f_\theta(\bar{d})} \;\; \forall\, \theta, \theta'.$$

That is, under MLRP, the space of possible observations is, indeed, ordered, and only kind of observation that matters is an observation of the best datapoint. There exist equilibria in which senders *only* send as great a mass as possible of $\bar{d}$, as well as equilibria in which senders send both that and up to a fraction $\frac{f_{\bar{\theta}}(d)}{f_{\underline{\theta}}(d)}$ as much data of observation d.[11]

Endogenous data acquisition

We have assumed so far that the distribution of μ is exogenous and identical for all senders. This captures some sources of variation in the data volume, such as invalid trials due to human error or dropouts. But the volume of data generated may also vary because of sender-specific differences in data-gathering ability – either different capacities (e.g. time constraints) or costs of obtaining more evidence.

The case of exogenous capacities is simple, and there is a one-to-one mapping between capacity constraints and distributions of attained data. The outcome of the game is unchanged if, instead of assuming that senders are randomly endowed with a measure μ of data following the outcome of trials, we suppose that each sender knows their capacity K for data collection prior to experiments, which is uncorrelated with the state. Then the sender's optimal strategy in the data-collection stage is to meet their capacity exactly (set $\mu = K$), the distribution of μ over the population is the same as the distribution of K, and the receiver draws identical conclusions.

[11] In Dzuida, $\frac{f_{\bar{\theta}}(d)}{f_{\underline{\theta}}(d)} = 0$ for all d, so there is a unique optimal strategy in her (selected) equilibrium.

It is more challenging to map costs of data acquisition to disclosure game outcomes. Nevertheless, it is trivially true that every distribution of data endowments can be founded on *some* cost structure, as cost functions

$$c(\mu) = \begin{cases} 0, & \mu \leq K \\ u_s(\mathbb{1}_{\theta_1}) + 1, & \mu > K \end{cases}$$

mimic capacities in that the (possibly weakly) optimal choice is $\mu = K$, and so any g can be imitated by a corresponding distribution over K among such cost functions. Conversely, for most reasonable distributions of cost functions over the population, there must exist g such that the equilibrium outcome of the augmented game with data acquisition is the lexicographically undominated outcome of the disclosure game in which g is the distribution of endowments.

References

Isaiah Andrews and Maximillian Kasy. Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794, 2019.

Elchanan Ben-Porath, Eddie Dekel, and Barton L. Lipman. Mechanisms with evidence: Commitment and robustness. *Econometrica*, 87(2):529–566, 2019.

Jeremy Bertomeu and Davide Cianciaruso. Verifiable disclosure. *Economic Theory*, 65(4): 1–34, 2018.

Ronald A. Dye. Disclosure of nonproprietary information. *Journal of Accounting Research*, 23(1):123–145, 1985.

Wioletta Dzuida. Strategic argumentation. *Journal of Economic Theory*, 146(4), 2011.

Joseph Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5(4):514–531, 1993.

Mike Felgenhauer and Elisabeth Schulte. Strategic private experimentation. *American Economic Journal: Microeconomics*, 6(4):74–105, 2014.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461, 1981.

Sergiu Hart, Ilan Kremer, and Motty Perry. Evidence games: Truth and commitment. *American Economic Review*, 107(3):690–713, 2017.

Woon-Oh Jung and Young K. Kwon. Disclosure when the market is unsure of information endowment of managers. *Journal of Accounting Research*, 26(1):123–145, 1988.

Steven Matthews, Masahiro Okuno-Fujiwara, and Andrew Postlewaite. Refining cheap-talk equilibria. *Journal of Economic Theory*, 55(2):247–273, 1991.

Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391, 1981.

Daniel Rappoport. Evidence and skepticism in verifiable disclosure games. *Working paper*, 2022.

Hyun Song Shin. News management and the value of firms. *RAND Journal of Economics*, 21(1):58–71, 1994.

Hyun Song Shin. Disclosures and asset returns. *Econometrica*, 71(1):105–133, 2003.

Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 2014.

A Appendix A: Construction of imitation equilibrium

B Continuous-data imitation equilibrium

Proof of Theorem 4.1 We construct $u_k(\mu_k)$ that is monotone increasing – this implies that it must be almost-everywhere differentiable. Since it is also continuous, it is completely determined by its derivative over the points at which the derivative exists. To avoid confusion, we focus on the left derivative of u_k, which we denote by u_k^- and, analogously to the top-down construction of the finite-data equilibrium, we construct the payoff function starting from the top down, starting from the frontier $v = \theta_J$.

It is helpful to define

$$A_j(\mu) = \{\theta_k : k \in \arg\max_k u_k(\mu/r_j(k))\}$$

to be the set of states that type μf_j finds it weakly optimal to target given $\hat{\mu}$. The range of $u_k(\mu_k)$ is $[0, \theta_k]$ since no type of higher state ever targets state θ_k, so payoffs to targeting θ_k cannot exceed θ_k itself.

Define

$$S(v) := \{\theta_k : \theta_k > v\}$$

to be the set of states that, if known, yield a payoff greater than v. Then $\hat{\mu}_k(v) < \infty$ iff $\theta_k \in S(v)$, and since play is supported on $\{\hat{\mu}_k(u_k(\mu/r_j(k)))f_k : \theta_k \in A_j(\mu)\}$ and $\sigma(\hat{\mu}_k(u)f_k|\hat{\mu}_k(u)f_k) = 1$, $S(v)$ is exactly the set of states that are targeted by some type under σ to obtain a payoff of v.

Given a burden of proof vector $\hat{\mu}(v) = (\hat{\mu}_k(v))_{\theta_k \in S(v)}$, the associated *frontier* consists of all types that are just able to meet some component of $\hat{\mu}(v)$, with no slack, that is, all types

$\tilde{\mu}_j f_j$ such that

$$r_j(k)\tilde{\mu}_j = \hat{\mu}_k(v) \text{ for some } \theta_k \in S(v), \text{ and } \nexists \theta_{k'} \in S(v) \text{ s.t. } r_j(k')\tilde{\mu}_j > \hat{\mu}_{k'}(v).$$

The frontier is then $\tilde{\mu}[\hat{\mu}(v)] = (\tilde{\mu}_1, \ldots, \tilde{\mu}_{l-1}, \hat{\mu}_l(v), \ldots, \hat{\mu}_J(v))$ if $S(v) = (\theta_l, \ldots, \hat{\theta}_J)$.

Let the set of states under which some type of sender obtains payoff v and finds it weakly optimal to target state θ_k be

$$\tau_{\hat{\mu}}^{opt}(\theta_k, v) = \{\mu f_j : \hat{\mu}_k(v) f_k \in \tilde{S}_j(\mu)\}$$

and let the set of states such that some type of sender obtains payoff v by targeting a state θ_k with strictly positive probability under σ be

$$\tau_{\hat{\mu}}^{supp}(\theta_k, v) = \{\mu f_j : \hat{\mu}_k(v) f_k \in \text{supp } \sigma(\cdot|\mu f_j)\}.$$

Of course, $\tau_{\hat{\mu}}^{supp}(\theta_k, v) \subseteq \tau_{\hat{\mu}}^{opt}(\theta_k, v)$.

For convenience of notation, we extend the definitions of these set-valued functions to any set of inputs (rather than a single input) by letting the function of the set be the union of the function applied to each individual element of the input set: thus for every set S of states, $\tau_{\hat{\mu}}^{opt}(S, v) = \bigcup_{\theta_k \in S} \tau_{\hat{\mu}}^{opt}(\theta_k, v)$ and $\tau_{\hat{\mu}}^{supp}(S, v) = \bigcup_{\theta_k \in S} \tau_{\hat{\mu}}^{supp}(\theta_k, v)$, and for every set ω of measures of data, $A_j(\omega) = \bigcup_{\mu \in \omega} A_j(\mu)$.

Additionally, we define the expectation of the state under the receiver's belief that the the sender is a type that receives v under $\hat{\mu}$ and finds it weakly optimal to target a state in S as follows.

$$V_{\hat{\mu}}(S, \hat{\mu}(v)) = \frac{\sum_{\theta_j \in \tau_{\hat{\mu}}^{supp}(S,v)} \beta_0(\theta_j)\theta_j g(\tilde{\mu}_j[\hat{\mu}(v)]) \frac{d\tilde{\mu}_j[\hat{\mu}(v)]}{dv}}{\sum_{\theta_j \in \tau_{\hat{\mu}}^{supp}(S,v)} \beta_0(\theta_j) g(\tilde{\mu}_j[\hat{\mu}(v)]) \frac{d\tilde{\mu}_j[\hat{\mu}(v)]}{dv}}.$$

In contrast, the expectation of the state under the receiver's true belief over θ conditional on knowing that the sender has sent some message that yields payoff v and targets a state in S is

$$W_{\hat{\mu}}(S, v|\sigma) := \frac{\sum_{\theta_j \in \tau_{\hat{\mu}}^{opt}(S,v)} \beta_0(\theta_j)\theta_j g(\tilde{\mu}_j[\hat{\mu}(v)]) \frac{d\tilde{\mu}_j[\hat{\mu}(v)]}{dv} \sigma(\{\hat{\mu}_k f_k\}_{\theta_k \in S} | \tilde{\mu}_j[\hat{\mu}(v)] f_j)}{\sum_{\theta_j \in \tau_{\hat{\mu}}^{opt}(S,v)} \beta_0(\theta_j) g(\tilde{\mu}_j[\hat{\mu}(v)]) \frac{d\tilde{\mu}_j[\hat{\mu}(v)]}{dv} \sigma(\{\hat{\mu}_k f_k\}_{\theta_k \in S} | \tilde{\mu}_j[\hat{\mu}(v)] f_j)} = v. \quad (1)$$

For any partial strategy $\hat{\sigma}$ that gives mixing probabilities between the messages $\hat{\mu}_{k_i}(v)\mathbf{f}$, the payoff $W_{\hat{\mu}}(S, v|\hat{\sigma}(v))$ is always weakly greater than $V_{\hat{\mu}}(S, v)$The two are equal exactly when all types obtaining payoff v that find it weakly optimal to target a state in M do so with probability 1.

Fix a frontier $\hat{\mu}(v)$, where $\theta_{l-1} < v \leq \theta_l$. It will be useful to define an undirected graph $H(v)$ on $S(v)$ by adding an edge between θ_k and $\theta_{k'}$ if and only if $\tau_{\hat{\mu}}^{opt}(\theta_k, v) \bigcap \tau_{\hat{\mu}}^{opt}(\theta_{k'}, v) \neq \emptyset$, that is, if there is some type that finds it optimal to target either state θ_k or state $\theta_{k'}$, and is indifferent between the two. Let C be the collection of connected components of $H(v)$.

We use the following algorithm to partition $S(v)$ at a given frontier $\hat{\mu}(v)$.

Algorithm: This algorithm calculates the payoffs to targeting a state in $S(v)$ at frontier $\hat{\mu}(v)$ when all types that do not obtain higher payoffs than v and who can target some $\hat{\mu}_k(v)f_k, \theta_k \in S(v)$ target the highest-payoff of these messages among those that they can, and assigns states θ_k to the same partition element if, across them, $\hat{\mu}_k(v)f_k$ must result in the same payoff, and for α close to 1, $\alpha\hat{\mu}_k(v)f_k$ must also result in the same payoff, so that for states under which types at the frontier are indifferent between such messages, they remain so for nearby frontiers.

First, note that if σ is such that, when there is a collection of states $\Sigma \subseteq S(v)$ such that, over an interval of payoffs, there always exists between any 2 states in Σ a path of other states in Σ such that there are types that mix with interior probability between any two successive states, then for all $\theta_k, \theta_{k'} \in \Sigma$,

$$\frac{r_j(k)}{r_j(k')} = \frac{\hat{\mu}_k(u)}{\hat{\mu}_{k'}(u)} = \frac{\frac{d\hat{\mu}_k(u)}{du}}{\frac{d\hat{\mu}_{k'}(u)}{du}} \quad \left(= \frac{\frac{du_k(\hat{\mu}_{k'}(u))}{d\mu}}{\frac{du_k(\hat{\mu}_k(u))}{d\mu}} \right) \tag{2}$$

for all u in the interval of payoffs and for all j that target some state in Σ at the frontier $\hat{\mu}(u)$.

We define

$$\Delta_n(\Sigma, \hat{\alpha}) := \frac{d^n}{d\alpha^n} \frac{\sum_{\theta_j \in \tau_{\hat{\mu}}^{supp}(\Sigma,v)} \beta_0(\theta_j)\theta_j g\left(\alpha\tilde{\mu}_j[\hat{\mu}(v)]\right) \tilde{\mu}_j[\hat{\mu}(v)]}{\sum_{\theta_j \in \tau_{\hat{\mu}}^{supp}(\Sigma,v)} \beta_0(\theta_j) g\left(\alpha\tilde{\mu}_j[\hat{\mu}(v)]\right) \tilde{\mu}_j[\hat{\mu}(v)]}\Bigg|_{\alpha=\hat{\alpha}}.$$

This is equal to the nth derivative of the payoff to the set of senders in states that target a state in Σ with positive probability at frontier $\hat{\mu}(v)$, that have an amount $\hat{\alpha}\tilde{\mu}_j[\hat{\mu}(v)]$ of data, when we assume that eq. 2 holds over Σ.

Start with a collection of assigned partition elements, $\mathcal{A}_0 = \emptyset$, and a collection of sets of unassigned states, $\mathcal{C}_0 = C$. Given $\mathcal{A}_n$ and $\mathcal{C}_n$, initialize $\mathcal{A}_{n+1} = \mathcal{C}_{n+1} = \emptyset$, and, taking each set $S \in \mathcal{C}_n$ sequentially, proceed as follows:

1. Take all subsets $\Sigma \subseteq S$ and calculate $\Delta_0(\Sigma, 1)$. Tiebreak any with the same value by $\Delta_1(\Sigma, 1), \Delta_2(\Sigma, 1), \ldots$, successively, and take the largest subset Σ that is maximal. Label it with $\tau_{\hat{\mu}}^{supp}(\Sigma, v)$, and add it to $\mathcal{A}_{n+1}$.

   Note that this implies that $\frac{du_k(\hat{\mu}_k(v))}{d\mu_k} = \frac{\Delta_1(\Sigma,1)}{\hat{\mu}_k(v)}$ when equation 2 holds for $\theta_k, \theta_{k'} \in \Sigma$ over $[v - \epsilon, v]$, $\epsilon > 0$.

2. Take $S \setminus \Sigma$, and let $C(S)$ be the collection of connected components of the graph on S constructed analogously to $H(v)$. Add $C(S)$ to C_{n+1} (i.e. augment C_{n+1} as the union of itself and $C(S)$).

3. Repeat on $\mathcal{A}_{n+1}$ and $\mathcal{C}_{n+1}$ until $\mathcal{C}_{n+1} = \emptyset$.

Putatively, if senders of type $\alpha\tilde{\mu}_j[\hat{\mu}(v)]f_j$ with $\theta_j \in \tau_{\hat{\mu}}^{supp}(v)$ pooled with each other, then payoffs are equal to

$$u_k(\alpha\hat{\mu}_k(v)) = v_\Sigma(\alpha, \hat{\mu}(v)) := \left.\frac{\sum_{\theta_j \in \tau_{\hat{\mu}}^{supp}(\Sigma,v)} \beta_0(\theta_j)\theta_j g\left(\alpha\tilde{\mu}_j[\hat{\mu}(v)]\right)\tilde{\mu}_j[\hat{\mu}(v)]}{\sum_{\theta_j \in \tau_{\hat{\mu}}^{supp}(\Sigma,v)} \beta_0(\theta_j) g\left(\alpha\tilde{\mu}_j[\hat{\mu}(v)]\right)\tilde{\mu}_j[\hat{\mu}(v)]}\right|_{\alpha=\hat{\alpha}},$$

and the burden-of-proof function for $\underline{v} \leq v$ is given by

$$\mu_\Sigma^{put}(\underline{v}) := \{v_\Sigma^{-1}(\underline{v}, \hat{\mu}(v))\hat{\mu}_k(v)f_k\}_{\theta_k \in S(v)},$$

where $v_\Sigma^{-1}(\underline{v}, \hat{\mu}(v))$ is the inverse of $v_\Sigma(\cdot, \hat{\mu}(v))$.

The reason that a partition element is a subset of targetable states in which all messages must achieve the same payoff at the is that, since Σ is a maximal highest-value subset over those that do not already have a higher value, it is either partitionable into smaller subsets, each of which also achieves the same value, or not; but in either case, in each minimal subset that achieves the maximal value, there is a path of messages between any two messages in the subset such that, in the targeting strategy, some type mixes with strictly positive probability between any two adjoining messages. The reason for this is that, for any smaller subset $\Sigma' \subset \hat{\Sigma}$, we have that $V_{\hat{\mu}}(\Sigma', \hat{\mu}(v)) < V_{\hat{\mu}}(\hat{\Sigma}, \hat{\mu}(v))$ if Σ is a minimal subset that achieves the maximal value. Since the expectation of the state conditional on knowing the message played is in $\hat{\Sigma}$ is at least $V_{\hat{\mu}}(\hat{\Sigma}, \hat{\mu}(v))$, there must be some message that yields payoff at least $V_{\hat{\mu}}(\Sigma, \hat{\mu}(v))$. But since there is no message, and indeed no proper subset of messages in $\hat{\Sigma}$ that achieve payoff $V_{\hat{\mu}}(\Sigma, \hat{\mu}(v))$ if all types that can play one of them do, it must be that for any subset, there is a type that can play some message in the subset but plays a message outside the subset with positive probability.

The reason the same holds true in frontiers to the left of $\hat{\mu}(v)$ is that, if $\Delta_0(\Sigma, 1)$ is uniquely maximal, then $\Delta_0(\Sigma, \alpha)$ is still greater than $\Delta_0(\Sigma', 1)$ for any Σ' and α sufficiently close to 1. So, in any state under which senders target a state in Σ at $\hat{\mu}(v)$, it remains optimal for them to do so for α close to 1, assuming the putative payoffs above. In addition, the putative payoffs are feasible, because every subset of Σ has lower value. If tiebroken by Δ_1, Δ_2, and so on, then although $\Delta_0(\Sigma, 1)$ is not uniquely maximal, Σ does maximize $\Delta(\cdot, 1)$ immediately to the left of $\hat{\mu}(v)$.

We will use the partition constructed by the algorithm to construct the equilibrium in chunks. For consistency, we want the following condition:

Condition 1. The value of each partition element constructed using the algorithm is the same, and is equal to v.

Under this condition, there is a partial strategy $\hat{\sigma}$ on each partition element such that $W_{\hat{\mu}}(\theta_k, v|\hat{\sigma}) = v$ for all states θ_k in the partition element, and furthermore, there is no partial strategy on a subset of messages in that partition element such that all messages in the subset result in the same payoff that is greater than v.

If Condition 1 holds at $\hat{\mu}(v)$ and Σ is the partition constructed using the algorithm at $\hat{\mu}(v)$, then there exists some $\epsilon > 0$ such that, for all $\underline{v} \in [v-\epsilon, v]$, Condition 1 holds for the frontier $\{v_\Sigma^{-1}(\underline{v}, \hat{\mu}(v))\hat{\mu}_k(v)f_k\}_{\theta_k \in S(v)}$. To show this, observe the following claim, which follows directly from statement of the condition and from continuity of $v_\Sigma(\alpha, \hat{\mu}(v))$:

Claim B.1 *Let the set of types that target a state in Σ and achieve a payoff of $\underline{v}$ under μ_Σ^{put} be $\tau_\Sigma^{put}(\underline{v})$.*

If Condition 1 holds at $\hat{\mu}(v)$, then if there exists no $v' \in (\underline{v}, v]$ such that either

1. *There is a type $t \in \tau_\Sigma^{put}(v')$ such that t can imitate a higher-value state, i.e. there exists partition element such that Σ' $v_{\Sigma'}^{-1}(v'', \hat{\mu}(v))\hat{\mu}_k(v)f_k \subseteq t$ for some $v'' > v'$*
2. *There is a partition element Σ with a subset $\Sigma' \subseteq \Sigma$ such that $v_{\Sigma'}(v_\Sigma^{-1}(v', \hat{\mu}(v)), \hat{\mu}(v)) > v'$,*

then Condition 1 continues to hold at $\hat{\underline{v}}$.

Note that, because for any partition element $\Sigma' \neq \Sigma$ either $v_{\Sigma'}^{-1}(v, \hat{\mu}(v))\hat{\mu}_k(v)f_k \not\subseteq t$, or $\Delta_n(\Sigma', 1) < \Delta_n(\Sigma, 1)$ for some n such that $\Delta_i(\Sigma', 1) = \Delta_i(\Sigma, 1)$ for all $i < n$, the continuity of $v_\Sigma(\alpha, \hat{\mu}(v))$ implies that for $\underline{v}$ close to v (1) cannot not hold. Again by continuity, (2) cannot hold for $\underline{v}$ close to v because for all $\Sigma' \subseteq \Sigma$, $v_{\Sigma'}(v_\Sigma^{-1}(v, \hat{\mu}(v)), \hat{\mu}(v)) \leq v$ and $\Delta_n(\Sigma', 1) < \Delta_n(\Sigma, 1)$ for some n such that $\Delta_i(\Sigma', 1) = \Delta_i(\Sigma, 1)$ for all $i < n$.

We will use this to construct the equilibrium in segments over which Condition 1 holds, and re-construct partitions using the algorithm in at most countably many points at which either (1) or (2) holds. Generically, the number of such points (and thus steps in the construction) is finite.

Now we turn to constructing larger pooling sets when there is a positive-measure set of types that can achieve the frontier payoff. Given that types support their play on $\{\hat{\mu}_k(u_k(\mu/r_j(k)))f_k : \theta_k \in A_j(\mu)\}$, and $u_k(\mu_k)$ is increasing, all types capable of sending a message in $\hat{\mu}(v) \cdot \mathbf{f}$ achieve a payoff of at least v. We define the set of types that are incapable of sending a message in $\hat{\mu}(v) \cdot \mathbf{f}$, but capable of sending a message in set M, as $T(v, M)$. We will denote the payoff to the sender of the receiver knowing they are one of a set of types that has positive probability measure under the receiver's prior as $U(T)$, and in particular,

$$U(T(v,M)) = \frac{\sum_{j=1}^{J} \beta_0(\theta_j)\theta_j \max(\max_{k \geq l}(G(\frac{\hat{\mu}_k(v)}{r_j(k)})) - \min\{G(\mu) : \exists m \in M \text{ s.t. } m \subseteq \mu f_j\}, 0)}{\sum_{j=1}^{J} \beta_0(\theta_j) \max(\max_{k \geq l}(G(\frac{\hat{\mu}_k(v)}{r_j(k)})) - \min\{G(\mu) : \exists m \in M \text{ s.t. } m \subseteq \mu f_j\}, 0)}.$$

Note that $\sup_M U(T(v, M)) \geq v$, because $\lim_{\alpha \to 1} U(T(v, \alpha\hat{\mu})) = v$. If there is a positive-measure type set $T(v, M)$ that achieves the value $\sup_M U(T(v, M))$, then take the largest such set and call it $\hat{T}^{max}_{\hat{\mu}}(v)$. Then the following hold:

1. If there exists a set $T(v, M)$ that achieves the value $\sup_M U(T(v, M))$, then there is a unique largest set that does so, and so $\hat{T}^{max}_{\hat{\mu}}(v)$ is well-defined.

2. Whenever $\hat{T}^{max}_{\hat{\mu}}(v)$ exists, there exist $\mu_l, \ldots, \mu_J$ such that $\hat{T}^{\max}_{\hat{\mu}}(v) = T(v, \{\mu_l f_l, \ldots, \mu_J f_J\})$.

3. Whenever $\hat{T}^{max}_{\hat{\mu}}(v)$ exists, there exists a partial strategy $\hat{\sigma} : \hat{T}^{\max}_{\hat{\mu}}(v) \to M = \{\mu_l f_l, \ldots, \mu_J f_J\}$ such that the payoff to any message $m \in M$ given that senders in $\hat{T}^{\max}_{\hat{\mu}}(v)$ play according to $\hat{\sigma}$ is $\hat{U}_{\hat{\mu}}(v)$.

The first point follows from the fact that, unless the union of two such sets yields payoff at least $\hat{U}_{\hat{\mu}}(v)$, then their intersection – which corresponds to the pool of types implemented by a different message set – yields strictly greater payoff. To see the 2nd point, simply take μ_k to be the minimum amount of data distributed f_k such that the dataset still contains a message in M, for each $k \geq l$, and note that the resulting set of types is a subset of $T(v, M)$ that has a smaller mass of types θ_j, $j < l$ but the same mass of types θ_k, $k \geq l$. Since $U(T(v, M)) \geq v \geq \theta_{l-1}$, this can only improve the payoff to the pool. The last point comes from the fact that, if $\hat{T}^{\max}_{\hat{\mu}}(v)$ is a maximum-payoff pool, then for each subset $S \subseteq M$, the payoff to the pool implemented by S is no greater than $U(\hat{T}^{\max}_{\hat{\mu}}(v))$, which is sufficient to ensure that $\hat{\sigma}$ exists. In addition, $U(T(v, M))$ is absolutely continuous with respect to every component of $\hat{\mu}(v)$ and each μ_k.

Lemma B.2 *If $\hat{T}^{\max}_{\hat{\mu}}(v)$ exists, then Condition 1 is satisfied by the burden of proof vector $M = \{\mu_l f_l, \ldots, \mu_J f_J\}$ such that $\hat{T}^{\max}_{\hat{\mu}}(v) = T(v, M)$.*

Proof Suppose not; then one of two cases is true:

1. <u>There is a collection of states $\Sigma \subset S(v)$ such that $V_{\hat{\mu}}(\Sigma, M) > v$.</u>

   Then, since $V_{\hat{\mu}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J)$ is continuous in α, there is $\underline{\alpha} < 1$ such that $V_{\hat{\mu}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J) > v$ for all $\alpha \in [\underline{\alpha}, 1]$. Consider an alternative type set, $T(M_{\underline{\alpha},\Sigma}, v)$ where $M_{\underline{\alpha},\Sigma}$ includes the messages $\mu_k f_k$ for $\theta_k \in S(v) \setminus \Sigma$, and the messages $\underline{\alpha}\mu_k f_k$ for $\theta_k \in \Sigma$.

   For $\underline{\alpha}$ small enough, the set of types in $T(M_{\underline{\alpha},\Sigma}, v) \setminus T(M, v)$ includes exactly those in frontiers $(\alpha M)^1_{\alpha=\underline{\alpha}}$ that find it weakly optimal to target a state in Σ. So, the expectation of the state given that the sender's type is in $T(M_{\underline{\alpha},\Sigma}, v) \setminus T(M, v)$ exceeds v, and so $T(M_{\underline{\alpha},\Sigma}, v)$ is higher-payoff than $T(M, v)$, contradicting that $T(M, v) = \hat{T}^{\max}_{\hat{\mu}}(v)$.

2. <u>There is a element of the partition, $\Sigma' \subset S(v)$, such that $V_{\hat{\mu}}(\Sigma', M) > v$.</u>

   Then WLOG let Σ' be the lowest-value element of the partition. Similarly to the above, since $V_{\hat{\mu}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J)$ is continuous in α, there is $\bar{\alpha} > 1$ such that $V_{\hat{\mu}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J) < v$ for all $\alpha \in [1, \bar{\alpha}]$. Consider an alternative type set, $T(M_{\bar{\alpha},\Sigma'}, v)$ where $M_{\bar{\alpha},\Sigma'}$ includes the messages $\mu_k f_k$ for $\theta_k \in S(v) \setminus \Sigma'$, and the messages $\bar{\alpha}\mu_k f_k$ for $\theta_k \in \Sigma'$.

For $\bar{\alpha}$ small enough, the set of types in $T(M,v)\setminus T(M_{\bar{\alpha}},\Sigma')$ includes exactly those in frontiers $(\alpha M)_{\alpha=1}^{\bar{\alpha}}$ that find it weakly optimal to target a state in Σ. Then the expectation of the state given that the sender's type is in $T(M,v)\setminus T(M_{\bar{\alpha}},\Sigma')$ is less than v, so the expectation given that the type is in $T(M_{\bar{\alpha}},\Sigma')$ exceeds v, contradicting that $T(M,v)=\hat{T}^{\max}_{\hat{\mu}}(v)$.

Since neither case is possible, M, taken as the payoff frontier corresponding to v, must satisfy Condition 1.

The iterative algorithm to construct the equilibrium of 4.1 starts from the highest-potential-payoff senders and creates payoff frontiers that satisfy Condition 1. It proceeds as follows:

1. Start with $l=J$ and $\hat{\mu}_J(\theta_J)=1$.
2. For each l, construct frontiers $\hat{\mu}_k(v)$ as follows:
   (a) Start at $v=\theta_l$ and burden-of-proof vector $\hat{\mu}(\theta_l)$, as constructed from the previous step. For all $v>\theta_l$, let $\hat{\mu}(v)$ be as already constructed. Define
   $$\check{\mu}_l(\theta_l)=\max\{\mu:\exists j<l \text{ s.t. } \tilde{\mu}_j[\hat{\mu}(\theta_l)]\geq\mu\},$$
   and rewrite $\hat{\mu}(\theta_l)=(\check{\mu}_l(\theta_l),\hat{\mu}_{l+1}(\theta_l),\ldots,\hat{\mu}_J(\theta_l))$. Proceed as below to rewrite $\hat{\mu}(v)$ for $v<\theta_l$:
   (b) Fix $S=\{\theta_k\}_{k=l}^J$. Given the frontier $\hat{\mu}(v)$, check if $\hat{T}^{max}_{\hat{\mu}}(v)$ exists, and if so, find $M=\{\mu_l f_l,\ldots,\mu_J f_J\}$ that implements $\hat{T}^{max}_{\hat{\mu}}(v)$ and rewrite $\hat{\mu}(v)=M$.
   (c) At $\hat{\mu}(v)$, using the algorithm, partition S into subsets of states, and calculate $v_\Sigma(\alpha,\hat{\mu}(v))$ for all $\alpha\in[0,1]$ for each subset. Take the lowest-value frontier, $\hat{\mu}(v')$, under putative payoffs $v_\Sigma(\alpha,\hat{\mu}(v))$ such that the conditions of Claim B.1 are satisfied and such that $\hat{T}^{max}_{\hat{\mu}}(v'')$ does not exist for any $v''\in(v',v]$, and assign strategies according to Algorithm 2 between $\hat{\mu}(v)$ and the new frontier $\hat{\mu}(v')$.
   (d) Set $v=v'$ and set $\hat{\mu}(v')$ as the new frontier, and repeat the above 2 steps until $v'=0$.
3. Repeat the above steps for each l in descending order until $l=1$, and fix the resulting $\hat{\mu}$.

It remains to show that the equilibrium constructed is the unique lexicographically undominated equilibrium. First we show that it is lexicographically undominated. To see this, suppose for the sake of contradiction that it is dominated by another equilibrium σ'. Let v be the lowest value such that payoff frontiers for all $v'>v$ are identical under σ and σ'.

There exists some $u\leq v$ such that $t^+_{\sigma'}(u)\setminus t^+_{\sigma^{im}}(u)$ is nonempty and $t^+_{\sigma^{im}}(u)\setminus t^+_{\sigma'}(u)$ is empty. Let $T_\sigma:=t^+_{\sigma^{im}}(u)\setminus t^+_{\sigma^{im}}(v)$, let $T_{\sigma'}:=t^+_{\sigma'}(u)\setminus t^+_{\sigma'}(v)$, and let $T:=t^+_{\sigma'}(u)\setminus t^+_{\sigma^{im}}(u)=T_{\sigma'}\setminus T_{\sigma^{im}}$. Note that $T_{\sigma'}=T_{\sigma^{im}}\sqcup T$ and that T has positive measure: for any $\mu f_j\in T$, we have $\tilde{\mu}_j[\hat{\mu}(u)]>\mu$, so for all $\mu'\in[\mu,\tilde{\mu}_j[\hat{\mu}(u)]]$, $\mu' f_j\in T$.

By the construction of σ^{im}, $\mathbb{E}[\theta|T] < u$, otherwise in step 2(c) there would be $\hat{T}^{max}_{\hat{\mu}}(u)$ with value equal to (at least) the frontier of u under σ'. By law of iterated expectations,

$$\begin{aligned}
\mathbb{E}[\theta|T_{\sigma'}] &= Pr(T_{\sigma^{im}}|T_{\sigma'})\mathbb{E}[\theta|T_{\sigma^{im}}] + Pr(T|T_{\sigma'})\mathbb{E}[\theta|T] \\
&= Pr(T_{\sigma^{im}}|T_{\sigma'})\mathbb{E}\left[\mathbb{E}_{\sigma^{im}}[\theta|m]\big|m \in \cup_{t\in T_{\sigma^{im}}}\text{supp}(\sigma^{im}(\cdot|t))\right] + Pr(T|T_{\sigma'})\mathbb{E}[\theta|T] \\
&< Pr(T_{\sigma^{im}}|T_{\sigma'})\left(\int_{t\in T_{\sigma^{im}}} \mathbb{E}[\theta|m \in \text{supp}(\sigma^{im}(\cdot|t))]q(t|t\in T_{\sigma^{im}})dt\right) + Pr(T|T_{\sigma'})u \\
&= Pr(T_{\sigma^{im}}|T_{\sigma'})\left(\int_{t\in T_{\sigma^{im}}} u_{\sigma^{im}}(t)q(t|t\in T_{\sigma^{im}})dt\right) + Pr(T|T_{\sigma'})u
\end{aligned} \tag{3}$$

where $q(t|t \in T_{\sigma^{im}})$ is the restriction of the prior probability density over types to $T_{\sigma^{im}}$.

Alternatively, we can write

$$\begin{aligned}
\mathbb{E}[\theta|T_{\sigma'}] &= Pr(T_{\sigma^{im}}|T_{\sigma'})\mathbb{E}[\mathbb{E}[\theta|t]|t \in T_{\sigma^{im}}] + Pr(T|T_{\sigma'})\mathbb{E}[\theta|T] \\
&= \int_{t\in T_{\sigma'}} \mathbb{E}_{\sigma'}[\theta|m \in \text{supp}(\sigma'(\cdot|t))]q(t|t\in T_{\sigma'})dt \\
&= Pr(T_{\sigma^{im}}|T_{\sigma'})\int_{t\in T_{\sigma^{im}}} \mathbb{E}_{\sigma'}[\theta|m \in \text{supp}(\sigma'(\cdot|t))]q(t|t\in T_{\sigma^{im}})dt \\
&\quad + Pr(T|T_{\sigma'})\int_{t\in T} \mathbb{E}_{\sigma'}[\theta|m \in \text{supp}(\sigma'(\cdot|t))]q(t|t\in T)dt \\
&= Pr(T_{\sigma^{im}}|T_{\sigma'})\int_{t\in T_{\sigma^{im}}} u_{\sigma^{im}}(t)q(t|t\in T_{\sigma^{im}})dt + Pr(T|T_{\sigma'})\int_{t\in T} u_{\sigma^{im}}(t)q(t|t\in T) \\
&\geq Pr(T_{\sigma^{im}}|T_{\sigma'})\int_{t\in T_{\sigma^{im}}} u_{\sigma^{im}}(t)q(t|t\in T_{\sigma^{im}})dt + Pr(T|T_{\sigma'})u.
\end{aligned} \tag{4}$$

Since both cannot hold, σ' cannot lexicographically dominate σ^{im}.

In addition, if σ' is any other equilibrium, then it is lexicographically dominated. Since σ' does not lexicographically dominate σ^{im}, there exists u such that $t^+_{\sigma^{im}}(u)\setminus t^+_{\sigma'}(u)$ is nonempty. Then, by Lemma B.3, if M is the set of messages that yield a payoff of at least u under σ', then the expectation of θ over the set of types $T = (\mathcal{T}_\infty \setminus \mathcal{T}^+_\infty(M)) \setminus \mathcal{T}^+_\infty(\hat{\mu}_{\sigma^{im}}(u))$ is at least u.

Using this, we construct an equilibrium that lexicographically dominates σ' as follows. There must exist a subset of messages $M'_0 \in \{\hat{\mu}_{\sigma^{im},k}(u)f_k\}$ such that there is a partial strategy over all types in T that can send a message in M'_0 such that all such messages yield the same payoff, which is at least u. Let T_0' be the set of such types. If the payoff is exactly u, stop. Otherwise there exists a value $u_1' > u$ such that the value of the set of types that is either in T_0', or achieves a payoff between u and u_1' under σ', is no greater than u_1' and greater than the next-highest payoff that is achieved by any type under σ'. This set of types is exactly the set of types that can send some message in M_0' or some message that yields a payoff of at least u under σ', and cannot obtain a payoff of at least u_1' under σ'. There is some subset of the union of messages in M_0' and those that yield a payoff of at least u under σ' such that the value of the set of types that do not achieve a payoff of at least u_1' under σ' that can send one of those messages is at least u_1', and there exists a partial strategy over those types among those messages such that all obtain the same payoff. Call the set of messages and the set of types M_1' and T_1', respectively, and again, stop if the value is between u_1' and the next-highest payoff achieved on-path under σ'. Continue iterating until the process terminates; then there will be a set of messages that implements a set of types that will obtain a payoff u_n' that is higher than they achieve in σ', while all types that can obtain a greater payoff u_n' under σ' continue to do so. Then, construct an equilibrium with the same frontiers as σ' down to u_n', that continues with the frontier to u_n' given by the new set of messages, and such that the continuation for values below u_n' is arbitrary.[12] This equilibrium lexicographically dominates σ'.

B.1 Proofs of convergence (Theorem 6.1)

Proof The proof of theorem 6.1 has 3 steps. First, we give a lemma establishing that, for any set of messages M, when the set of all types in $\mathcal{T}_\infty \setminus \mathcal{T}_\infty^+(M)$ that attain a payoff of at least v in σ^{im} is nonempty, their payoff when they form a pool is at least v. Using this, we show that $u_{\sigma^{im}}$ is a lower bound on payoffs for types in $\mathcal{T}_\infty$ in the limit. Finally, Bayes plausibility implies that

$$\lim_{N\to\infty} \sum_{j=1}^{j} \beta_0(\theta_j) \int_{\mu=0}^{1} u_{\sigma_N}(\mu f_j) g(\mu) d\mu = \mathbb{E}_{q_\infty}[\mathbb{E}_{\beta(\cdot|\sigma(t))}[\theta]|t] = \mathbb{E}_{\beta_0}[\theta],$$

which in conjunction with the lower bound implies that in the limit outcomes must coincide exactly with $u_{\sigma^{im}}$ for types in $\mathcal{T}_\infty$.

Lemma B.3 *If M is a collection of messages and $(\underline{\mu}_1 f_1, \ldots, \underline{\mu}_i f_i; \underline{\mu}_{i+1} f_{i+1}, \ldots, \underline{\mu}_J f_J)$ is the frontier of types achieving a payoff of at least v under σ^{im}, where $\theta_i < v \le \theta_{i+1}$, then*

$$\mathbb{E}[\theta | t \in \mathcal{T}_\infty^+(\{\underline{\mu}_j f_j\}_{j=1}^J) \setminus \mathcal{T}_\infty^+(M)] \ge v$$

whenever $\mathcal{T}_\infty^+(\{\underline{\mu}_j f_j\}_{j=1}^J) \setminus \mathcal{T}_\infty^+(M)$ is nonempty.

[12] It is possible that all possible continuations below u_n', fixing the frontier u_n' as the one constructed, involve creating some pool with value greater than u_n'. But in this case, simply re-iron by adding any types above the frontier capable of sending a message in the higher-payoff pool to the higher-payoff pool until it is no longer higher-payoff.

Proof of Lemma Denote $T(v,M) = \mathcal{T}^+_\infty(\{\underline{\mu}_j f_j\}_{j=1}^J)\backslash\mathcal{T}^+_\infty(M)$. Let $(\bar\mu_1,\dots,\bar\mu_i;\bar\mu_{i+1},\dots,\bar\mu_J)$ be the minimum masses of data distributed like $f_1,\dots,f_i;f_{i+1},\dots,f_J$, respectively, necessary to send some message in M. Then

$$\mathbb{E}[\theta|t\in T(v,M)] = \frac{\sum_{j=1}^J \beta_0(\theta_j)\theta_j(G(\bar\mu_j)-G(\underline{\mu}_j))}{\sum_{j=1}^J \beta_0(\theta_j)(G(\bar\mu_j)-G(\underline{\mu}_j))}.$$

If $(\bar\mu_{i+1},\dots,\bar\mu_J)\le(\underline{\mu}_{i+1},\dots,\underline{\mu}_J)$ pointwise, then $T(v,M)$ is empty. Otherwise, let the states $j_1,\dots,j_A$ be the maximal set such that $(\bar\mu_{j_1},\dots,\bar\mu_{j_A})>(\underline{\mu}_{j_1},\dots,\underline{\mu}_{j_A})$ pointwise. Call the set of types that send $\mu' f_{j_a}$ with positive probability under σ^{im} by $\tau^{supp}_{\sigma^{im}}(\mu' f_{j_a})$, and let $\theta(t)$ refer to the state corresponding to the distribution of dataset t. Denote by $\hat\sigma_v$ the partial strategy in which types, restricted to $T(v,M)$, play as they do in σ^{im}.

In the case when payoffs under $u_{\sigma^*_\infty}$ are strictly increasing at $\mu' f_{j_a}$,

$$\begin{aligned}\mathbb{E}_{\hat\sigma_v}[v(\theta)|\mu' f_{j_a}] &= \frac{\sum_{t\in\tau^{supp}_{\sigma^{im}}(\mu' f_{j_a})\bigcap T(v,M)}\theta(t)g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}}{\sum_{t\in\tau^{supp}_{\sigma^{im}}(\mu' f_{j_a})\bigcap T(v,M)}g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}}\\ &\ge \frac{\sum_{t\in\tau^{supp}_{\sigma^{im}}(\mu' f_{j_a})}\theta(t)g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}}{\sum_{t\in\tau^{supp}_{\sigma^{im}}(\mu' f_{j_a}))}g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}}\\ &\ge v.\end{aligned} \tag{5}$$

where the first inequality comes from the fact that $\theta_{j_a}\ge v>\theta(t)$ whenever $\theta(t)\ne\theta_{j_a}$, and $\mu' f_{j_a}\in T(v,M)$ only if all types that play it under σ^* are also in $T(v,M)$.

In the case where there are positive-measure sets T of senders achieving the same payoff $v'>v$ under σ^{im} with $T\bigcap T(v,M)$ nonempty, then let M' be the set of messages that implements the pool, and

$$\mathbb{E}_{\hat\sigma_v}[v(\theta)|m\in M'] = \mathbb{E}_{\hat\sigma_v}[v(\theta)|t\in T\bigcap T(v,M)].$$

The value of $T\setminus T(v,M)$ is equal to the value of $T\bigcup\mathcal{T}^+_\infty(M)$, which is no more than v' since $T=\hat T^{max}_{\hat\mu}(v')$, and so it contains no subsets of higher value. Therefore, $\mathbb{E}_{\hat\sigma_v}[v(\theta)|t\in T\bigcap T(v,M)]\ge v'\ge v$.

Then, taking the total expectation over both cases, the expectation of θ given that the sender's type is in $T(v,M)$ is a weighted average of $\mathbb{E}_{\hat\sigma_v}[\theta|\mu' f_{j_a}]$ over on-path messages $\mu' f_{j_a}$ in $T(v,M)$ in which the payoff is strictly decreasing; and the value over positive-measure sets of equal payoff. We have shown that each component is no less than v, and so the weighted average is also at least v. ∎

Before proceeding to construct bounds on payoffs in the finite games, it is helpful to define a neighborhood of $\mathcal{T}_\infty$ as the set of types in each finite game with datasets distributed similarly to the underlying distribution in some state. For $\eta \in (0,1]$ and $k \in [0,1]$, define

$$S_N(\eta,k) = \{t \in \mathcal{T}_N : |t| \geq k \text{ and } \exists\theta \text{ s.t. } \sup_d |t(d) - |t|f_\theta(d)| \leq \eta\}.$$

Fix an integer n. Conditional on $|t| = n$ and the true state being θ_j, the Glivenko-Cantelli theorem states that there is a bound on the probability that $\sup_d |\sum_{x=1}^d t(d) - \frac{n}{N}F_j(d)| > \eta$ that decreases to 0 for large n, irrespective of N. Because data have a discrete distribution, this implies a similar bound on the empirical probability mass function: if $|t| = n$ and θ_j is the true state, the probability that $\sup_d |t(d) - \frac{n}{N}f_j(d)| > \eta$ is at most $b_=(n,\eta)$, with $\lim_{n\to\infty} b_=(n,\eta) = 0$ for all $\eta > 0$. If the true state is $\theta_{j'} \neq \theta_j$ and $|t| = n$, then the probability that $\sup_d |t(d) - \frac{n}{N}f_j(d)| > \eta$ is at least $b_{\neq}(n,\eta)$, with $\lim_{\eta\to 0}\lim_{n\to\infty} b_{\neq}(n,\eta) = 1$.

When N and k are large, the proportion of types that lie in $S_N(\eta,k)$ is close to 1, for all η. In particular, $\lim_{k\to 0}\lim_{\eta\to 0}\lim_{N\to\infty} q_N(S_N(\eta,k)) = 1$, since:

- With probability decreasing to 0 as $k \to 0$, $|t| < k$.
- For fixed k and η, the probability that there does not exist θ such that $\sup_d |t(d) - |t|f_\theta(d)| \leq \eta$ given that $|t| \geq k$ decreases to 0 as $Nk \to \infty$.

We may further subdivide $S_N(\eta,k)$ into a set of types associated with each state,

$$S_N^j(\eta,k) = \{t \in S_N(\eta,k) : \sup_d |t(d) - |t|f_j(d)| \leq \eta\}.$$

A further consequence of the convergence of empirical distributions is that, when $Nk \to \infty$ and $\eta \to 0$, the sets $(S_N^j(\eta,k))_{j=1}^J$ are disjoint. Additionally, for all $t \in S_N^j(\eta,k)$, there is a uniform lower bound on the probability that the state is θ_j given that the sender is of type t, which we call $w(k,\eta,N)$, with $\lim_{k\to 0}\lim_{\eta\to 0}\lim_{N\to\infty} w(k,\eta,N) = 1$.

In addition, we can lower-bound $q_N(\{t \in S_N^j(\eta,k) : \underline{\mu}f_j \subseteq t \subseteq \bar{\mu}f_j\})$ for all $k < \underline{\mu} < \bar{\mu}$. Let $\Delta(N)$ be a bound on $\sup_d |(\sum_{x=1}^d g_N(x)) - G(d)|$ that goes to 0 as $N \to \infty$. Observe that if $\underline{\mu} + \eta < |t| < \bar{\mu} - \eta$ and $t \in S_N^j(\eta,k)$, then $\underline{\mu}f_j \subseteq t \subseteq \bar{\mu}f_j$, so a lower bound is

$$q_N(\{t \in S_N^j(\eta,k) : \underline{\mu}f_j \subseteq t \subseteq \bar{\mu}f_j\}) \geq \beta_0(\theta_j)(1 - b_=(Nk,\eta))(G(\bar{\mu} - D\eta) - G(\underline{\mu} + D\eta) - \Delta(N)). \tag{6}$$

Similarly, there is an upper bound on $q_N(\{t \in S_N^j(\eta,k) : t \not\subseteq \underline{\mu}f_j \text{ and } \bar{\mu}f_j \not\subseteq t\})$:

$$q_N(\{t \in S_N^j(\eta,k) : t \not\subseteq \underline{\mu}f_j \text{ and } \bar{\mu}f_j \not\subseteq t\}) \leq \beta_0(\theta_j)(G(\bar{\mu}+D\eta) - G(\underline{\mu}-D\eta) + \Delta(N)) + (1-\beta_0(\theta_j))b_{\neq}(kN,\eta). \tag{7}$$

Now we proceed to construct a lower bound for $u_{\sigma_N}(\hat{\mu}f_{\hat{j}})$. First, recall that $u_{\sigma_N}(\mu f_j) \geq \max_{\{f \in \mathcal{T}_N : t \subseteq \mu f_j\}} u_{\sigma_N}(t)$. Observe that there exists a dataset $\hat{t} = \frac{1}{N}(\lfloor N\hat{\mu}f_{\hat{\theta}}(1)\rfloor, \ldots, \lfloor N\hat{\mu}f_{\hat{\theta}}(k)\rfloor)$ in $\mathcal{T}_N$ and that $u_{\sigma_N}(\hat{\mu}f_{\hat{j}}) \geq u_{\sigma_N}(\hat{t})$.

For a given N, suppose $\hat{t}$ belongs to the mth upper pool under the algorithm that constructs σ_N. Denote by $\hat{M}_N(m-1)$ the set of messages that implement the upper pools in step $1, \ldots, m-1$, and fix $\mathcal{T}_{N,m} = \mathcal{T}_N^+(\hat{M}_N(m-1))$ to be the set of remaining types at the start of the mth step of the algorithm that constructs σ_N; therefore, $\hat{t}$ belongs to $\mathcal{T}_{N,m}$.

Let $\underline{M}_\infty(\epsilon, N)$ be the set of on-path messages that result in a payoff of $u_\sigma((\hat{\mu}-\epsilon)f_{\hat{j}})$ under infinite data. We see that the set of types in $\mathcal{T}^+_{N,m}(\underline{M}_\infty(\epsilon, N))$ includes $\hat{f}$ when N is large enough. From Lemma B.3, there is an upper pool in $\mathcal{T}_{\hat{N},m}$ that achieves a payoff of at least $u(\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon, N)))$, so $u_{\sigma_N}(\hat{\mu}f_{\hat{j}})$ is lower-bounded by $u(\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon, N)))$.

Let $(\underline{\mu}_1(\epsilon, N), \ldots, \underline{\mu}_J(\epsilon, N))$ be a vector that gives the minimum mass of data under distributions $f_1, \ldots, f_J$, respectively, such that the dataset contains some message in $\underline{M}_\infty(\epsilon, N)$, and let $(\bar{\mu}_1(N), \ldots, \bar{\mu}_J(N))$ be the maximum mass of data under each distribution such that there does not exist $t \in \mathcal{T}_{\hat{N},m}$ such that $t \subseteq \bar{\mu}_j f_j$. All $t \in \mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))$ satisfy $t \not\subseteq \underline{\mu}_j(\epsilon, N) f_j$ and $\bar{\mu}_j(N) f_j \not\subseteq t$, and all t satisfying $\underline{\mu}_j(\epsilon, N) f_j \subseteq t \subseteq \bar{\mu}_j(N) f_j$ for some j are in $\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))$.

We may rewrite

$$u(\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))) = \frac{\sum_{j=1}^J \sum_{t \in \mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))} q_N(t)\theta_j \pi_N(\theta_j|t)}{\sum_{t \in \mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))} q_N(t)}. \tag{8}$$

Let the numerator be $Q(N, \hat{\mu}f_{\hat{j}}, \epsilon)$ and the denominator be $R(N, \hat{\mu}f_{\hat{j}}, \epsilon)$. Analogously to eq. 6, a lower bound for $Q(N, \hat{\mu}f_{\hat{j}}, \epsilon)$ is

$$\underline{Q}(N, \hat{\mu}f_{\hat{j}}, \epsilon) = \sum_j \beta_0(\theta_j)\theta_j[G(\bar{\mu}_j(N)-\eta D) - G(\max(\underline{\mu}_j(\epsilon, N)+\eta D, k)) - \Delta(N)]w(k,\eta,N)(1-b_=(k,\eta)), \tag{9}$$

and it follows from eq. 7 that an upper bound for R is

$$\begin{aligned}\bar{R}(N, \hat{\mu}f_{\hat{j}}, \epsilon) = &\left(\sum_j \beta_0(\theta_j)[G(\bar{\mu}_j(N)+\eta D) - G(\underline{\mu}_j(\epsilon, N)-\eta D) + \Delta(N)]\right) \\ &+ J(1-b_{\neq}(k,\eta)) + (1 - q_N(S_N(\eta, k))).\end{aligned} \tag{10}$$

We have

$$\lim_{k\to 0}\lim_{\eta\to 0}\lim \inf_{N\to\infty} \underline{Q} \geq \lim \inf_{N\to\infty} \sum_{j=1}^J \beta_0(\theta_j)\theta_j(G(\bar{\mu}_j(N)) - G(\underline{\mu}_j(\epsilon, N)))$$

and

$$\lim_{k\to 0}\lim_{\eta\to 0}\lim \inf_{N\to\infty} \bar{R} \leq \lim \inf_{N\to\infty} \sum_{j=1}^J \beta_0(\theta_j)(G(\bar{\mu}_j(N)) - G(\underline{\mu}_j(\epsilon, N))).$$

Both of the RHS are finite and strictly positive for all N and $\epsilon > 0$; therefore,

$$\begin{aligned}\lim_{k\to 0}\lim_{\eta\to 0}\lim\inf_{N\to\infty}\frac{\bar{Q}}{\bar{R}} &\geq \liminf_{N}\frac{\sum_{j=1}^{J}\beta_0(\theta_j)\theta_j(G(\bar{\mu}_j(N)) - G(\underline{\mu}_j(\epsilon,N)))}{\sum_{j=1}^{J}\beta_0(\theta_j)(G(\bar{\mu}_j(N)) - G(\underline{\mu}_j(\epsilon,N)))} \\ &= \liminf_{N}\mathbb{E}[\theta|t\in T(u_{\sigma_\infty}((\hat{\mu}-\epsilon)f_{\hat{j}}),\hat{M}_N(m-1))] \\ &\geq u_{\sigma_\infty}((\hat{\mu}-\epsilon)f_{\hat{j}}),\end{aligned} \tag{11}$$

where the last inequality follows from Lemma B.3.

Because k and η are arbitrary variables used to obtain the bound, it follows from this that $\lim_{N\to\infty} u(\mathcal{T}^{+}_{\hat{N},m}(\underline{M}_\infty(\epsilon))) \geq u_{\sigma_\infty}((\hat{\mu}-\epsilon)f_{\hat{j}})$. Finally, because payoffs are continuous, taking a sequence of bounds as $\epsilon \to 0$ implies that $\liminf_{N\to\infty} u_{\sigma_N}(\hat{\mu}f_{\hat{j}}) \geq \lim_{\epsilon\to 0}\liminf_{N\to\infty} u(\mathcal{T}^{+}_{\hat{N},m}(\underline{M}_\infty(\epsilon))) \geq u_{\sigma_\infty}(\hat{\mu}f_{\hat{j}})$.

The last step is to show that

$$\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{1} u_{\sigma_N}(\mu f_j)g(\mu)d\mu = \mathbb{E}_{\beta_0}[\theta].$$

Since we know already that

$$\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{1} u_{\sigma_\infty}(\mu f_j)g(\mu)d\mu = \mathbb{E}_{\beta_0}[\theta]$$

and $\liminf_{N\to\infty} u_{\sigma_N}(\mu f_j) \geq u_{\sigma_\infty}(\mu f_j)$ for all $\mu f_j \in \mathcal{T}_\infty$, this additional fact suffices to ensure that $u_{\sigma_N}(\cdot) = u_{\sigma_\infty}(\cdot)$ over $\mathcal{T}_\infty$.

The proof comes from dividing $\mu \in (k,1)$ into X chunks, with the xth chunk given by $(\mu_{x-1},\mu_x]$ where $\mu_x = x\frac{1-k}{X} + k$.

Consider types $t \in S^j_N(\eta,k)$ such that $\mu_{x-1}f_j \subseteq t \subseteq \mu_x f_j$: their payoff under σ_N has to be in $[u_{\sigma_N}(\mu_{x-1}f_j), u_{\sigma_N}(\mu_x f_j)]$. This implies that

$$\begin{aligned}\underline{V}_N(k,\eta,X) &= \sum_{j=1}^{J}\beta_0(\theta_j)\sum_{x=1}^{X} u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1}-\eta D) - G(\mu_x+\eta D) - \Delta(N)](1-b_=(k,\eta)) \\ &\leq \mathbb{E}_{\beta_0}[\theta],\end{aligned} \tag{12}$$

since $\underline{V}_N(k,\eta,X)$ is a lower bound for the total probability-weighted sum of payoffs under σ_N over $t \in \mathcal{T}_N \bigcup S_N(\eta,k)$, while $\mathbb{E}_{\beta_0}[\theta]$ is equal to the total probability-weighted sum of payoffs under σ_N of all types in $\mathcal{T}_N$.

Finally, the difference between $\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{1}u_{\sigma_N}(\mu f_j)g(\mu)d\mu$ and $\underline{V}_N(k,\eta,X)$ vanishes as $X\to\infty$, $k\to 0$, $\eta\to 0$, and $N\to\infty$. To see this, observe that if c is an upper bound on g (which exists because g is continuous on compact interval $[0,1]$),

$$\begin{aligned} V_N(k,\eta,X) \geq \sum_{j=1}^{J}\beta_0(\theta_j)\Big(\sum_{x=1}^{X}&u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1})-G(\mu_x)] \\ &-(\theta_J b_=(k,\eta)[G(\mu_{x+1})-G(\mu_x)]+2c\eta D+\Delta(N))\Big) \\ \geq \sum_{j=1}^{J}\beta_0(\theta_j)&\sum_{x=1}^{X}u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1})-G(\mu_x)] \\ &-JX\theta_J(b_=(k,\eta)+2c\eta D+\Delta(N)). \end{aligned} \tag{13}$$

Then, for any ϵ and j, define $\xi_N^j(\epsilon,X)$ to be the set of values of x such that $u_{\sigma_N}(\mu_{x+1}f_j)-u_{\sigma_N}(\mu_x f_j)>\epsilon$. The size of $\xi_N^j(\epsilon,X)$ is at most $\frac{\theta_J}{\epsilon}$. For all $x\notin\xi_N^j(\epsilon,X)$, we have the bound $\int_{\mu_x}^{\mu_{x+1}}u_{\sigma_N}(\mu f_j)g(\mu)d\mu-u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1})-G(\mu_x)]<\epsilon[G(\mu_{x+1})-G(\mu_x)]$. So,

$$\begin{aligned} &\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{2}u_{\sigma_N}(\mu f_j)g(\mu)d\mu-\underline{V}_N(k,\eta,X) \\ &\leq\left(\sum_{j=1}^{J}\beta_0(\theta_j)\sum_{x=1}^{X}\left(\int_{\mu_x}^{\mu_{x+1}}u_{\sigma_N}(\mu f_j)g(\mu)d\mu-u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1})-G(\mu_x)]\right)\right) \\ &\quad+JX\theta_J(b_=(k,\eta)+2c\eta D+\Delta(N)+(1-q_N(S_N(\eta,k)))) \\ &\leq\sum_{j=1}^{J}\left(\beta_0(\theta_j)\left(\sum_{x\notin\xi_N^j(\epsilon,X)}\epsilon[G(\mu_{x+1})-G(\mu_x)]\right)+\left(\sum_{x\in\xi_N^j(\epsilon,X)}\theta_J[G(\mu_{x+1})-G(\mu_x)]\right)\right) \\ &\quad+JX\theta_J(b_=(k,\eta)+2c\eta D+\Delta(N)+(1-q_N(S_N(\eta,k)))) \\ &\leq\epsilon+J\frac{c(1-k)}{X}\frac{\theta_J^2}{\epsilon}+JX\theta_J\Big(b_=(k,\eta)+2c\eta D+\Delta(N)+(1-q_N(S_N(\eta,k)))\Big) \end{aligned} \tag{14}$$

since $\sum_{x\in\xi_N^j(\epsilon,X)}[G(\mu_{x+1})-G(\mu_x)]\leq\frac{c(1-k)}{X}\frac{\theta_J}{\epsilon}$. Then

$$\lim_{\epsilon\to0}\lim_{X\to\infty}\lim_{k\to0}\lim_{\eta\to0}\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{2}u_{\sigma_N}(\mu f_j)g(\mu)d\mu-\underline{V}_N(k,\eta,X)=\lim_{\epsilon\to0}\lim_{X\to\infty}\epsilon+J\frac{c(1-k)}{X}\frac{\theta_J^2}{\epsilon}=0.$$

Again, since ϵ, X, k, and η were all constructed variables, this implies that

$$\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{2}u_{\sigma_N}(\mu f_j)g(\mu)d\mu=\lim_{\epsilon\to0}\lim_{X\to\infty}\lim_{k\to0}\lim_{\eta\to0}\lim_{N\to\infty}\underline{V}_N(k,\eta,X)\leq\mathbb{E}_{\beta_0}[\theta].$$

As it is already clear from the lower bound on $u_{\sigma_N}(\mu f_j)$ that $\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{2}u_{\sigma_N}(\mu f_j)g(\mu)d\mu \geq \mathbb{E}_{\beta_0}[\theta]$, equality obtains.

B.2 Proof of Lemmas B.3, 5.2 and 5.4

Proof of Lemma B.3 Consider a game in which the type set is $\mathcal{T}^+(M)$, each type's action set is the set of messages in M that they are able to send, and the payoff to playing $\hat{\sigma}(\cdot|t)$ against the receiver's putative strategy profile $\hat{\sigma}'$ is $\sum_{\tilde{f}} u(\beta_{\hat{\sigma}'}(\cdot|\tilde{f}))\hat{\sigma}(\tilde{f}|t)$, the utility to the sender of the receiver's updated belief conditional on seeing them play $\tilde{f}$ when the population is expected to play according to $\hat{\sigma}'$.

Payoffs are continuous in $\hat{\sigma}$ and $\hat{\sigma}'$. Let the best response correspondence be given by

$$r_t(\hat{\sigma}') = \arg\max_{\hat{\sigma}(\cdot|t)} \sum_{\tilde{f}\in M} u(\beta_{\hat{\sigma}'}(\cdot|\tilde{f}))\hat{\sigma}(\tilde{f}|t).$$

A fixed point $\hat{\sigma}^*$ of r corresponds to a PBE of the constructed game, and a standard Nash existence argument shows that there must be at least one. Then let M' be the set of messages that achieve the highest payoff under $\hat{\sigma}^*$; along with the restriction of $\hat{\sigma}^*$ to $T^+(M')$, it forms an upper pool.

If M is not itself an upper pool, then $\mathcal{T}^+(M)\setminus\mathcal{T}^+(M')$ is nonempty and contains types that do worse than those in $\mathcal{T}^+(M')$. Then,

$$u(\mathcal{T}^+(M')) > u(\mathcal{T}^+(M)) > u(\mathcal{T}^+(M)\setminus\mathcal{T}^+(M)).$$

∎

Proof of Lemma 5.2 Consider 2 such pools, $M = \{\tilde{f}_1,\ldots,\tilde{f}_I\}$ and $M' = \{\tilde{f}'_1,\ldots,\tilde{f}'_J\}$, with type sets $\mathcal{T}^+(M)$ and $\mathcal{T}^+(M')$. We aim to show their union is also a utility-maximizing upper pool. Let $A = \mathcal{T}^+(M)\setminus\mathcal{T}^+(M')$, $B = \mathcal{T}^+(M')\setminus\mathcal{T}^+(M)$, and $C = \mathcal{T}^+(M)\bigcap\mathcal{T}^+(M')$. Observe that if we let M'' be the message set that includes $\tilde{f}_i \vee \tilde{f}'_j$ for every $i \leq I$, $j \leq J$ (where $\vee$ is the pointwise max operator on datasets), then $C = \mathcal{T}^+(\mathcal{M}'')$.

We have $u(\mathcal{T}^+(M)) = u(\alpha A + (1-\alpha)C) = u(\mathcal{T}^+(M')) = u(\alpha' B + (1-\alpha')C) = u^*$. So $u(\mathcal{T}^+(M)\bigcup\mathcal{T}^+(M')) \geq u^*$ unless $u(A) < u^*$, $u(B) < u^*$, and $u(C) > u^*$; but by the previous lemma, the last of these would imply that C contains a higher-utility upper pool than M and M'. Since this is not true, $u(\mathcal{T}^+(M)) = u^*$ and it is an upper pool itself (otherwise it would contain a strictly better upper pool, a contradiction). ∎

Proof of Lemma 5.4 Suppose to the contrary that $u(\mathcal{T}^+_m(M_m)) \leq u(\mathcal{T}^+_{m+1}(M_{m+1}))$. Then,

$$u(\mathcal{T}^+_m(M_m\bigcup M_{m+1})) = u(\alpha'\beta(\cdot|\mathcal{T}^+_m(M_m)) + (1-\alpha')\beta(\cdot|\mathcal{T}^+_{m+1}(M_{m+1}))) \geq u(\mathcal{T}^+_m(M_m)),$$

which implies (by Lemma B.3) that either $M_m\bigcup M_{m+1}$ must itself be an upper pool with respect to $\mathcal{T}_m$, or that there exists $M' \subset M_m\bigcup M_{m+1}$ such that $u(\mathcal{T}^+_m(M')) > u(\mathcal{T}^+_m(M_m))$. Either of these would contradict that M_m is a maximal upper pool in $\mathcal{T}_m$.

B.3 Proof of Claim 3.2

Proof We proceed inductively. Since $\pi(0,N) > \pi(n_1,n_2)$ for all (n_1,n_2), and $(0,N)$ cannot be imitated by any other type, $M_1 = (0,N)$ and $T_{\hat{\sigma}_{M_1}} = \{(0,N)\}$.

Now suppose $M_m = (0, \tilde{n}_2[m])$ for $m = 1,\dots,j$. Then $\mathcal{T}_{j+1} = \{(n_1,n_2)\}_{n_2 \le \tilde{n}_2[j]-1, n_1 \le N-n_2}$. Consider any $(\tilde{n}_1,\tilde{n}_2)$. Then by combinatorial identity,

$$\beta(H|\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2)) = \pi_H(\tilde{n}_1,\tilde{n}_2).$$

There may be a set of types who are able to send $(\tilde{n}_1,\tilde{n}_2)$ but are already included in $T_{\hat{\sigma}_{M_m}}$ for some $m \le j$. This set, $\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2) \setminus \mathcal{T}_{j+1}$, satisfies

$$\beta(H|\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2) \setminus \mathcal{T}_{j+1}) = \beta(H|\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2[j])) = \pi_H(\tilde{n}_1,\tilde{n}_2[j]) > \pi_H(\tilde{n}_1,\tilde{n}_2).$$

Therefore,

$$\beta(H|\mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)) < \pi_H(\tilde{n}_1,\tilde{n}_2).$$

This implies that a single message $(\tilde{n}_1,\tilde{n}_2)$ with $\tilde{n}_1 > 0$ cannot be a highest-payoff pool, since the type set of the upper pool consisting of message $(0,\tilde{n}_2)$ yields strictly higher payoff. To see this, observe that

$$\beta(H|\mathcal{T}^+_{j+1}(0,\tilde{n}_2)) = \alpha\beta(H|\mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)) + (1-\alpha)\beta(H|\mathcal{T}^+_{j+1}(0,\tilde{n}_2) \setminus \mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)),$$

and $\beta(H|\mathcal{T}^+_{j+1}(0,\tilde{n}_2) \setminus \mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)) > \pi_H(\tilde{n}_1,\tilde{n}_2)$: that is, the receiver's belief conditional on the sender being in $\mathcal{T}_{j+1}$ and being able to send $\tilde{n}_2$ high signals but not $\tilde{n}_1$ low signals, is better than their belief when the sender is able to send at least $\tilde{n}_1$ low signals, therefore their belief is better when the burden of proof does not require any low signals be sent.

In addition, the highest-payoff pool cannot correspond to a set of distinct messages $M = \{(\tilde{n}_1^1,\tilde{n}_2^1),\dots,(\tilde{n}_1^L,\tilde{n}_2^L)\}$, such that $n_2^l < n_2^{l-1}$ and $n_1^l > n_1^{l-1}$ for all l.[13] To see this, first focus on

$$T_L := \{(n_1,n_2) : \tilde{n}_2^L \le n_2 < \tilde{n}_2^{L-1}, \tilde{n}_1^L \le n_1 \le N - n_2\},$$

the set of types in $\mathcal{T}_{j+1}$ that can send $(\tilde{n}_1^L,\tilde{n}_2^L)$ but no other messages in M. Observe as before that $\beta(H|T_L) \le \pi_H(\tilde{n}_1^L,\tilde{n}_2^L)$. Consider 2 cases:

- If $\beta(H|T_L) \ge \beta(H|\mathcal{T}_{j+1}(M))$, then let $M' = \{(\tilde{n}_1^1,\tilde{n}_2^1),\dots,(\tilde{n}_1^{L-2},\tilde{n}_2^{L-2}),(\tilde{n}_1^{L-1},\tilde{n}_2^L)\}$, i.e. replace messages $(\tilde{n}_1^{L-1},\tilde{n}_2^{L-1})$ and $(\tilde{n}_1^L,\tilde{n}_2^L)$ with a single message that is their (pointwise) minimum.
- If $\beta(H|T_L) < \beta(H|\mathcal{T}_{j+1}(M))$, then letting $M' = \{(\tilde{n}_1^1,\tilde{n}_2^1),\dots,(\tilde{n}_1^{L-1},\tilde{n}_2^{L-1})\}$, i.e. drop $(\tilde{n}_1^L,\tilde{n}_2^L)$ from the message set.

[13] A set of messages that does not satisfy these properties can either be reordered to do so, or is redundant in that there are some l, l' such that $(n_1^l,n_2^l) \subset (n_1^{l'},n_2^{l'})$; so sets of messages satisfying these criteria are exhaustive of possible upper pools.

In either case, we have $\beta(H|\mathcal{T}_{j+1}(M')) \geq \beta(H|\mathcal{T}_{j+1}(M))$, and M' is a strictly smaller set of messages than M. Repeat on M' and iterate until the message set is a singleton; then it is a commuting upper pool that yields strictly better belief than M.

The above argument shows that message set M_{j+1} of the unique upper pool chosen in the $j+1^{st}$ step of the algorithm is of the form $\{(0, \tilde{n}_2[j+1])\}$ where $\tilde{n}_2[j+1] < \tilde{n}_2[j]$. It is immediate the value of $\tilde{n}_2[j+1]$ that maximizes payoff to the pool is as given in the claim. Given the choice of $\tilde{n}_2[j+1]$, the payoff to M_m is decreasing in m; therefore, the strategy profile constructed is an equilibrium. ∎

B.4 Relationship to Hart et al. (2017)

Truth-leaning equilibrium, Hart et al. (2017) An equilibrium σ is truth-leaning if, whenever $t \in \arg\max_{m \subseteq t} u_s(\beta_\sigma(\cdot|m))$, then $\sigma(t|t) = 1$, and if $\sigma(t|t') = 0$ for all t', then $\beta_\sigma(\cdot|t) = \mathbb{1}_t$.

Lemma B.4 *Equilibria are truth-leaning if and only if they are immune to credible inclusive announcements.*

Proof An equilibrium is truth-leaning only if in it, any sender that does not tell the truth obtains a strictly better payoff than they would if their type was known. If in a truth-leaning equilibrium σ there is a credible inclusive announcement attaining payoff v for type set $T = t_1, \ldots, t_L$, either the type t_l plays t_l in equilibrium, or the payoff to the belief $\pi(t_l)$ is worse than the payoff that t_l obtains in equilibrium. If t_l plays t_l then T also includes all other types that play t_l in equilibrium. Since the payoff to either $\pi(t_l)$ for t_l not played in equilibrium, or the payoff to knowing the sender is one of the types that plays t_l in equilibrium, must both be no more than v, and for at least some t_l it must be strictly less, the payoff to knowing that the sender's type is in T must also be less than v, a contradiction. Thus, the unique truth-leaning equilibrium is also the unique equilibrium robust to credible inclusive announcements.