Essays on Communication and Signaling: Evidence, Inference, and Persuasion

by

Ying Gao

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by: Ying Gao
Department of Economics
May 15, 2024

Certified by: Drew Fudenberg
Paul A. Samuelson Professor of Economics, Thesis Supervisor

Certified by: Stephen Morris
Peter A. Diamond Professor of Economics, Thesis Supervisor

Accepted by: Isaiah Andrews
Professor of Economics
Departmental Committee on Graduate Studies

Essays on Communication and Signaling: Evidence, Inference, and Persuasion

by

Ying Gao

Submitted to the Department of Economics
on May 15, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

This thesis contains 3 chapters, each of which explores the implications of partial disclosure in a different context. As disclosures affect the beliefs of key actors in each scenario, optimal information design is an important consideration, and I try to characterize it and identify potential pitfalls in each case.

Chapter 1 considers the disclosure problem of a sender with a large dataset of hard evidence. A sender may have an incentive to drop observations before submitting the data to receivers in order to persuade them to take a favorable action. I predict which observations the sender discloses using a model with a continuum of data, and show that this model approximates the outcomes with large, multi-variable datasets. In the receiver's preferred equilibrium, the sender's strategy relies on imitation: they submit evidence that imitates the natural distribution under a more desirable target state. As a result, it is enough for an experiment to record data on outcomes that maximally distinguish higher states. I characterize these strategies and show that senders with little data or a favorable state fully disclose their data, but still suffer from the receiver's skepticism, and therefore are worse-off than they are under full information. On the other hand, senders with large datasets can benefit from voluntary disclosure by dropping observations under low states.

In Chapter 2, my coauthors[1] and I study the Federal Reserve's problem of disclosing the models it uses in supervisory stress tests of large banks. Banks argue that nondisclosure leads to inefficiencies stemming from uncertainty, but regulators are concerned that full disclosure can lead to banks gaming the system. We formalize the intuition behind this trade-off in a stylized model where both the regulator and banks have imperfect, private "models" about a risky asset, and the regulator uses its own model to 'stress test' the investment. We show that if the regulator uses its model to test the banks' investment, full disclosure is suboptimal, and the regulator may benefit from hiding the model when the bank's model

[1]My coauthors are Marc de la Barrera and Bumsoo Kim. We each contributed to the core results and writing of this paper; I would like to thank Marc for initiating a discussion of the Fed's stress testing policies, and Bumsoo for his considerable contribution to and encouragement on this project as part of his 2nd year paper.

is more precise than the regulator's own model. The key idea is that hiding the regulator's model forces the bank to guess it using the bank's own models, effectively eliciting the bank's private information. We also show that if the regulator can fine-tune disclosure policies, the regulator can approximately enforce banks to take the first-best action, through an intuition closely related to the Cremer and McLean (1988) information rent extraction result.

Finally, in Chapter 3 I return to signaling via hard evidence, and analyze communication through a piece of verifiable evidence when the receiver/decision maker is uncertain about where the sender's preferred action lies in relation to their own, that is, the sender's relative bias. In contrast to the known-preference case, fully informative communication is impossible, even when the receiver is certain the sender is informed. The main novelty is that receivers cannot distinguish between senders of opposite preferences who pool by withholding their information when it is unfavorable. Two opposing patterns of partial disclosure emerge. When senders are biased relative to receivers, but just as state-sensitive, nondisclosure is driven by senders with extreme preferences, who choose to withhold slightly unfavorable evidence. The reverse, however, occurs when senders are state-insensitive.

JEL Codes: D80, D82, G28

Thesis supervisor: Drew Fudenberg

Title: Paul A. Samuelson Professor of Economics

Thesis supervisor: Stephen Morris

Title: Peter A. Diamond Professor of Economics

Acknowledgments

I would like to thank more people than I can name for their contributions to my research and my personal well being during the writing of this thesis and my PhD. First among them are my advisors, who have been incredible sources of insight and support, and to whom I owe a great debt.

Drew has, at every step, been an unerring source of guidance, encouragement, and inspiration. Before I had even the inkling of an idea, he always encouraged me to consider the possibilities of new questions. By collaborating with him, I have learned from example about how to tackle hard problems, and this has changed my way of thinking very much for the better. His generous feedback and incisive thoughts on the framing of ideas has improved these papers enormously, and he has inspired an interest in many topics that I hope yet to pursue. I will miss the frequency of our meetings and, moreover, his enthusiasm and compassion.

Stephen's thoughtful suggestions and positivity have been enormously helpful, and he was a great source of advice when I was stuck in the research process. He has always shown me that there are insights where I might not have thought to seek them. Our discussions have never failed to boost my morale.

I must also mention my coauthors, including Marc and Bumsoo, with whom I wrote the 2nd chapter of this thesis. Being able to coauthor with friends and classmates was one of the highlights of my time in graduate school, for which I am very grateful. In addition, I would like to thank my other coauthors – the very insightful Harry, Nicole, Brendan, Markus, Whitney, Jerry, and Ben, who took a chance in working with a graduate student.

A number of other people contributed through their comments to work in this thesis. In particular, I thank Bob Gibbons and Alex Wolitzky for their extensive help with individual chapters, and feedback that ended up guiding the work that followed.

Thanks also to the MIT economics department, my cohort, and all the others I met during my time here for the lovely environment, and to my friends, especially Jacob, for supporting me and making these times to look back on.

Finally, I would not be who I am without my family — I must thank my mother for her care and thoughtfulness, and for being my role model in making decisions; my father, for his

optimism, his support in hard times, and the curiosity he has always shared with me; and my brother Steven and cousin Leo, for their honesty, understanding, and companionship.

Contents

List of Figures

List of Tables

Chapter 1

Inference from Selectively Disclosed Data

1.1 Introduction

Many decisions – including technology adoption, regulatory approval, and research grant-making – are based on self-disclosed data. The datasets used can often be very large, on the order of tens of thousands of trials for drug approval, and often hundreds of thousands of datapoints about locations and sales in merger cases. In and of themselves, big datasets may paint an accurate picture of reality, but this becomes less clear if the sender can disclose them strategically. Oftentimes, it is easier to verify that submitted data are real than that they are complete, in which case deciding which observations are admissible to include in the dataset is largely at the sender's discretion, even in the presence of mandatory disclosure norms.

We want to understand the role that disclosed data play in strategic communication between the sender and the decision-maker when receivers have uncertainty about the underlying dataset from which the sender extracted the submitted data. We consider the case of a sender with state-independent motives to persuade the receiver towards a particular action, and a receiver who observes a dataset the sender discloses, but interprets it with partial skepticism that the data are incomplete. Equilibrium play between the sender and receiver involves the sender submitting data as "proof" that the receiver should take a favorable action, and the receiver evaluating how persuasive the proof is depending on how likely it is sent by a sender with less persuasive data who has trimmed some discouraging observations. This can be modeled under the framework of an evidence game in which senders that have access to datasets with weakly more observations of each outcome can always mimic senders with fewer observations. A special case, in which senders either have or do not have access to a single data point, with probability known to the receiver, is already well-understood (Dye 1985), and demonstrates that senders can manipulate the receiver by disclosing nothing when the evidence is sufficiently poor.

Our primary innovation is to characterize disclosure in the opposite extreme, when datasets contain many observations. We propose a continuous-data model of the asymptotic distribution over potential datasets of the sender that depends on two things: the true state of the world that generates the data, and a random variable that describes the amount of data the sender collects. The continuum assumption captures the fact that empirical distributions are approximately deterministic in the limit with large numbers, and allows us to eliminate uncertainty over the randomness of draws, which makes the model more tractable than directly modeling large, finite N. Instead, we show that the outcome we characterize in the continuous model describes the limit outcome of communication in finite-data games as $N \to \infty$.

In addition to an extensive list of observations, a second characteristic feature of "big" data is a large outcome space. This motivates the novel use of a framework that encompasses general statistical settings, including those in which outcome and state spaces are large and the relationship between them complex. In particular, we place essentially no restrictions on the state-contingent data distribution. In general, unlike in a "good news-bad news" model of data, the ranking of states and the shape of their data-generating distributions endogenously affects the interpretation of different outcomes, and context determines whether data take on a positive or negative connotation.

Indeed, our first main result, Prop. 1, is a sufficiency result that says that in the receiver-optimal partial pooling equilibrium outcome, the state-contingent experimental outcome distribution affects the information transmitted only through a handful of key features: what matters are the observations of outcomes with the greatest likelihood ratio under a better vs. a worse state. Strikingly, since the distribution of data that distinguishes one state from another depends solely on the relative probabilities of likelihood ratio-maximizing outcomes, a receiver who wants to distinguish a relatively small number of states with many observations of high-dimensional data can do just as well restricting the dataset to only retain information about these outcomes. When state-contingent distributions of experimental outcomes satisfy the monotone likelihood ratio property (MLRP), we return to the case of only one "good news" outcome, and distinguishing it from other outcomes is sufficient to support receiver-optimal communication.

Our second result, Theorem 2, characterizes an "imitation" equilibrium implementation of the receiver-optimal equilibrium outcome, in which senders always show the receiver a dataset that can correspond to a naturally-generated dataset, so that on path, the receiver always places positive probability on the event that the sender is sending all their data. However, the receiver also infers from some datasets that the sender has with positive probability observed data corresponding to a different state than the revealed data suggest, but has dropped observations in order to imitate a more favorable distribution. When MLRP fails, it is important for the imitating sender to send a large-enough mass of realizations of a certain outcome, but not too much. The resulting outcome benefits senders under low states with more data at the expense of senders with less data in high states, since the former pool with the latter. The extent of pooling depends on the receiver's uncertainty about the

sender's data collection capabilities: the greater the variance in the receiver's belief about how much data the sender starts out with, the more senders can profitably imitate other senders, with outcomes converging to the full-information one as uncertainty vanishes.

In section 1.3.1, we provide an algorithm to accompany the characterization that constructs the limit game equilibrium outcomes using a top-down logic: senders with more data receive weakly greater payoffs, and we can construct the payoff frontiers of the continuous payoff function by specifying the burden of proof, or how much data of a given state's distribution a sender needs, to induce a particular belief in the receiver. The algorithm is applicable to any number of states and to any datasets with finite support, and we illustrate it with representative 2 and 3 state examples.

Finally, we turn to governing interactions in which the sender can resort to strategic omission, and show that two intuitive classes of information interventions can be effective: experimental design that tailors the set of observed outcomes to the set of states that the receiver wishes to distinguish, and public information provision to inform the receiver about ex-ante opaque aspects of the data generating process. The former has implications for research and testing guidelines, while the latter offers a formal channel through which decision-making depends on transparency in research practices.

1.1.1 Related literature

Strategic disclosure has been studied since the work of Grossman (1981) and Milgrom (1981), which showed that full disclosure is the unique outcome when receivers know that the sender wishes to prove the value of a good is high using verifiable information that they could choose to disclose. The assumption that receivers know the sender is informed is crucial to this benchmark, as Dye (1985) and Jung and Kwon (1988) show. They consider a case in which the sender has access to a single, real-valued piece of evidence with interior probability $p \in (0, 1)$, and shows that only senders for whom the evidence exceeds a threshold will choose to disclose it, with the rest withholding it in order to pool with those senders who lack evidence altogether. Shin (1994, 2003) shows that in the case where senders have an uncertain endowment of good news and bad news, the fact that senders withhold bad enough evidence implies a "sanitation equilibrium", in which all bad news is disposed of.

We extend these results by considering evidence structures with large, multidimensonal datasets. In our data-based setting, evidence is neither exogenously good nor bad, but the receiver draws inferences statistically, based on knowledge of the relationship between relevant state-parameters and the distributions of data they generate. The setting we consider encompasses the settings above, and captures a special case of more abstract evidence games of the type considered by Green and Laffont (1986), Okuno-Fujiwara et al. (1990), Hart et al. (2017), and Glazer and Rubinstein (2006). The main focus in those settings has been on eliciting good outcomes for the receiver with receiver-optimal mechanisms and equilibria with full revelation, and we follow in this spirit by fully characterizing a receiver-optimal partial pooling equilibrium in the game we study. Hart et al. (2017) in particular is foundational to

our equilibrium selection criterion. Their observation that the optimal deterministic mechanism, the receiver-optimal equilibrium, and the unique truth-leaning equilibrium all yield the same outcome generalizes straightforwardly to our setting.[1]

Rappoport (2022) and Jiang (2022) use an iterative algorithm to solve for truth-leaning equilibrium outcomes in finite evidence games, but it is computationally demanding to use it in games with large type spaces, and therefore infeasible to directly compute the large-N limit of outcomes in the strategic disclosure of finite datasets. Our approach is instead to use a continuous-data approximation to solve for asymptotic outcomes without explicitly computing outcomes of finite-data games, and to show that it is exactly the big-data limit outcome.

Only one other paper that we know of, by Dzuida (2011), uses a continuous measure of evidence to solve for communication with verifiable evidence. Unlike our model, hers considers evidence with a continuum of states but a binary "good news-bad news" outcome structure, and assumes there is a positive likelihood of a behavioral, honest type of the sender. The existence of the honest type, along with imposing continuity of payoffs in reported outcomes, selects an equilibrium resembling the truth-leaning one. As in our paper, providing interior amounts of negative evidence can be optimal in otherwise sanitation-like equilibria. We show that in the absence of honest types but with large outcome spaces, this can come about for a novel reason – sending interior amounts of observations of some outcomes can be necessary to simultaneously imitate desirable types while ruling out undesirable ones, and can be strictly preferable to sending less or more data.

We also relate to a broader literature about the optimal collection and disclosure of evidence, that considers costly (Migrow and Severinov (2022)) and dynamic evidence acquisition (Felgenhauer and Schulte 2014, Henry and Ottaviani 2019), sender-optimal disclosure mechanisms (Haghtalab et al. 2022), and discretionary disclosure after test or information design (Shiskin 2022, Dasgupta et al. 2022). Several papers use a restricted notion of evidence but are also explicitly concerned with the effect of allowing sample selection: Fishman and Haggerty (1990) and Di Tillio et al. (2021) study the case in which only a subset of observations are disclosed, and give conditions under which it is better that an informant have discretion over which data are selected.

Finally, there is a small literature of empirical findings about common patterns of voluntary disclosure. Some work in econometrics by Simonsohn et al. (2014), Andrews and Kasy (2019) and others studies the bias that arises from a range of exogenous patterns of selective reporting, and describes inference procedures that correct for it. In addition, a small body of experimental work studies how subjects disclose evidence when incentivized to persuade receivers in the lab. Jin et al. (2021) finds evidence that receivers' inferences are

[1]The optimal mechanism equivalence result has been noted by others, including Glazer and Rubinstein (2006), Sher (2011), and Ben-Porath et al. (2019), who show that the fact that commitment is not necessary for the optimum is robust to other settings, in particular with binary actions and multiple senders with type-dependent preferences.

often biased by accounting insufficiently for the sender's nondisclosure, and Li and Schipper (2020) shows that senders are also biased towards naive, truthful behavior. Both suggest that these behaviors are consistent with an initial lack of higher-order sophistication that is remedied, to some extent, by experience. On the other hand, Osun and Ozbay (2021) suggest that in a binary-type evidence game, senders' disclosure policies and receivers' commitment policies differ from those predicted by Hart et al. (2017) in ways consistent with a common understanding that senders are inherently averse to lying.

1.2 Model

States and payoffs. There is a sender (S), who wishes to communicate to a receiver (R) about an unknown state of the world, $\theta \in \Theta = \{\theta_1, \ldots, \theta_J\}$. The sender and receiver share a common prior $\beta_0(\cdot)$ over Θ. We assume that the receiver takes an action $a_r \in \mathbb{R}$ and that $\theta_1, \ldots, \theta_J$ are real numbers ordered with $\theta_j \leq \theta_{j+1}$, representing the optimal action for the receiver under each state, if it was known with certainty. The sender's payoff is simply (a monotone function of) a_r;[2] in short, regardless of their type, they want to induce the receiver to take the highest action possible.

Finally, we assume the receiver has an expected utility that is differentiable and single-peaked at the action that matches their expectation of the value of θ, that is, that for any belief $\beta \in \Delta\Theta$, the receiver's expected payoff $\mathbb{E}_\beta[u_r(a)]$ is single-peaked at $a_r(\beta) = \mathbb{E}_\beta[\theta]$.[3] We work with the sender's indirect utility as a function of the receiver's beliefs, which induces them to maximize the receiver's posterior expectation of θ:

$$u_s(\beta) = \mathbb{E}_\beta[\theta]. \tag{1.1}$$

For example, when the receiver is a policymaker, states can represent the true optimal policy. While the policymaker might be uncertain, they wish to enact a policy that matches the optimal policy in expectation, while the sender wishes for them to take as high an action as possible.

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 of observations drawn from a finite set of outcomes, $\mathcal{D} = \{1, \ldots, D\}$. The underlying data-generating distribution is state-contingent: under state θ_j, the observations are i.i.d. draws from distribution f_j.

We model the amount of data the sender has access to as a mass, $\mu \in [0, 1]$, that represents the fraction of total potential data that the sender can access, and has a continuous

[2] Because the receiver will always play a pure strategy, the sender's problem is unchanged if their payoffs are rescaled through a monotone mapping.

[3] The assumption of single-peakedness is necessary to identify the receiver-optimal equilibrium and the receiver-optimal mechanism.

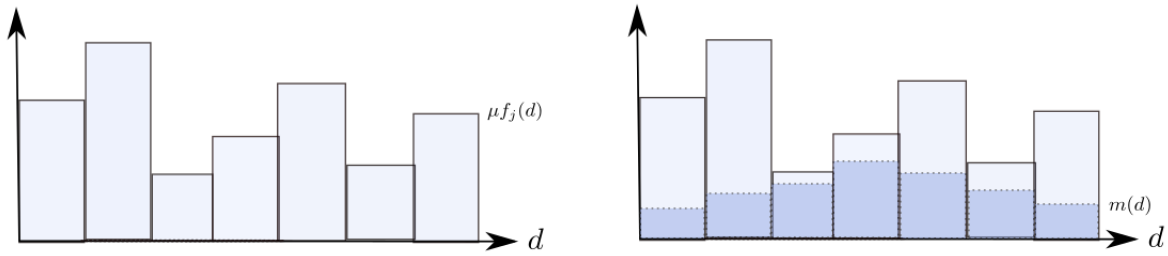


Figure 1.1: A feasible type and a feasible message.

distribution, g, that is state-independent[4], supported on $[0,1]$ with $g(1) = 0$, and infinitely left-differentiable[5]. The continuum assumption models big datasets in which the large number of draws essentially removes all uncertainty about the impact of randomly realized outcomes on the sender's dataset: conditional on state θ_j, the empirical distribution of data the sender observes is certain to be f_j, and μ does not affect the distribution of their evidence, only the amount of it. In other words, with probability 1, a sender with a mass μ of data under state θ_j observes the dataset $t = \mu f_j$. Any nonzero measure of data fully informs the sender of the state, and the set of possible complete datasets and types of the sender is $\mathcal{T} = [0,1] \times \Theta$.

The receiver, on the other hand, is uninformed about how much data the sender has. Their prior belief about the sender's type is given by the density

$$q(\mu f_j) = \beta_0(\theta_j) g(\mu). \tag{1.2}$$

Messaging and inference. Senders can choose a subset of observations from their dataset to submit to the receiver. We assume total flexibility in the choice of subset:

Assumption 1. *The sender can send any message $m \in \mathcal{M} = [0,1] \times \Delta\mathcal{D}$ that is a* subset *of their dataset ($m \tilde{\subseteq} \mu f_j$), where*

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

That is, a sender can drop an arbitrary mass of observations from their data, and then show the remaining ones to the receiver. By dropping observations, they can arbitrarily alter the relative frequencies of each outcome in the submitted dataset in order to imitate

[4]For simplicity of exposition, we focus on the case in which their belief about μ conditional on θ is given by a probability density g that is independent of θ, although most results hold identically for cases in which the distribution of μ is state-specific.

[5]The assumption that g has a vanishing right tail ensures that it is continuous on $\mathbb{R}^+$ while being supported on $[0,1]$, and simplifies the equilibrium construction: specifically, it ensures that the equilibrium payoffs are continuous in μ.

any distribution. However, this is costly in that it reduces the size of the submitted dataset, which is observable.

We have that $\mathcal{M} \supset \mathcal{T}$: the message space contains the set of all possible complete datasets, but also a D-dimensional set of other datasets that could be disclosed to the receiver after excluding part of their dataset. For any set of messages M, define the upper set $U(M)$ to be the set of types that can send a message in M, and for any set of types T, define the lower set $L(T)$ as the set of messages that some $t \in T$ can send.

Call the disclosure game with these parameters $\mathcal{G}(\Theta, \mathcal{D}, \beta_0, \{f_j\}_{j=1}^J, G)$. Upon observing the sender's message, the receiver updates their belief about the sender's type to $q(t|m)$, and then forms a new belief about the state,

$$\beta(\theta_j|m) = \frac{\sum_{j=1}^J \int_{\mu=0}^1 q(\mu f_j|m)\theta_j}{\sum_{j=1}^J \int_{\mu=0}^1 q(\mu f_j|m)}. \tag{1.3}$$

1.2.1 Equilibrium

The sender plays a messaging strategy $\sigma^* : \mathcal{T} \to \Delta\mathcal{M}$, knowing which the receiver infers the content of message they receive. As usual, the equilibrium we consider will be a Perfect Bayesian Equilibrium (Fudenberg and Tirole 1970), that is, $\beta^*(\cdot|m)$ must be consistent with the sender's strategy σ^*, and the sender must optimize, so $\sigma^*(m|t) > 0$ only if $m \in \arg\max_{m' \in L(t)} \mathbb{E}_{\beta(\cdot|m')}[\theta]$.

Call the map from types to payoffs, $u_{\sigma^*}(t)$, the *outcome* of the equilibrium.[6] In the perfectly separating outcome, the sender obtains a payoff of θ. As in Milgrom (1981), Grossman (1981), and Dye (1985), when g is a degenerate distribution such that μ is known to the receiver, then all attempts to mislead the receiver unravel, and the fully separating outcome obtains in every PBE. When g and all f_j have full support, there is partial pooling in every PBE. However, PBE are often not unique, and in this case, there may be multiple β^*, differing on off-path messages, that are consistent with σ^*, and the game generically has multiple, non-payoff-equivalent PBE outcomes. Any message that can be played by some type of sender under a state $\theta_j \geq \mathbb{E}_{\beta_0}[\theta]$ is played on-path in some PBE.

Intuition suggests that the game is fundamentally one of imitation: senders tailor their data to increase the receiver's belief that the state is a higher one, and they can only do so by imitating the datasets submitted by higher-state types, who themselves may be imitating others or trying to distinguish themselves as well as possible from lower-state types. One way to imitate a higher-state type of sender is to try to prove you have all the data that they would, and no more – that is, to imitate their complete dataset. We define an *imitation equilibrium* to capture the idea that sender masquerades as other type by imitating their full datasets.

[6] This is a departure from the usual definition of an outcome of an extensive-form game, but consistent with the definition in Hart et al. (2017) and Rappoport (2022). It describes the action the receiver plays after communicating with each type, and so describes the consequences of communication in the game.

Definition 1.2.1. (σ^*, β^*) *is an imitation equilibrium if it is an equilibrium, and under* σ^**,*

a. Every on-path message is in $\mathcal{T}$*,*

b. Type μf_j *plays* $m \neq \mu f_j$ *if and only if* $\theta_j < \max_{m' \in L(t)} \mathbb{E}_{\beta^*(\cdot|m')}[\theta]$*, and otherwise reports their full dataset.*

In other words, with an imitation messaging strategy every type of the sender either fully reveals their data or imitates another type's full dataset, and they only consider the latter if it could give them a better payoff than letting the receiver be fully informed of the state.

Why do we focus on these equilibria? Imitation equilibria are *truth-leaning*, as first defined by Hart et al. (2017) in the context of general evidence games with finite types. The idea applies identically in this setting. Formally, given a base game $\mathcal{G}$, for $\epsilon = (\epsilon_t, \epsilon_{t|t})_{t \in \mathcal{T}}$, let a game $\mathcal{G}_\epsilon$ be the game with an identical type set and type distribution, but with two differences. First, type t's payoffs to playing t are perturbed by ϵ_t, so that t's payoff to playing t is $\mathbb{E}_{\beta(\cdot|t)}[\theta] + \epsilon_t$. Secondly, type t plays t with at least probability $\epsilon_{t|t}$ – i.e. with probability $\epsilon_{t|t}$ a sender with dataset t is a commitment type that plays their full dataset regardless of whether doing so is optimal, while with probability $1 - \epsilon_{t|t}$ type t is strategic. A truth-leaning equilibrium is an equilibrium of the base game that can be obtained as a limit of equilibria of ϵ-perturbed games as $\epsilon \to 0$.

While truth-leaning equilibrium strategies capture a sender's slight bias towards truth-telling, the truth-leaning equilibrium outcome has desirable properties in its own right. When the receiver's expected payoffs are single-peaked in their action, the truth-leaning equilibrium outcome is also receiver-optimal, and is the outcome of the optimal mechanism when the receiver can commit to a single action as a response to each message. This is well-known in the finite case studied by Hart et al. (2017), and continues to be true in the continuous model that we study. It is also the only equilibrium outcome robust to a slightly stronger version of a credible announcement (Matthews et al. 1991). We say that under equilibrium σ^* a collection T of types of the sender can benefit from an *inclusive* credible announcement if there is a set of messages such that T is comprised of every type that 1) finds some message in the set feasible and 2) weakly benefits from the receiver updating that their type is in T from the prior, relative to the receiver's equilibrium inference; and there is at least one type in T that strictly benefits.[7] Robustness to such announcements means that the equilibrium survives even if senders are able to override the receiver's beliefs by proposing sensible reinterpretations of messages, and can coordinate to do so; it rules out, for example, play that is "stuck" in a bad equilibrium due to immalleable off-path beliefs. For a deeper discussion of these refinements, see Hart et al. (2017) and Appendix A.2.

Claim 1. *Imitation equilibrium messaging strategies are the truth-leaning equilibrium messaging strategies of* $\mathcal{G}$*. Imitation equilibrium outcomes are:*

[7] The departure from the usual credible announcement is that the collection of types making the announcment must also contain all types who are indifferent between participating in the announcement and their equilibrium payoff.

- *Receiver-optimal among equilibria and deterministic mechanisms;*
- *The unique inclusive announcement-proof equilibrium outcomes.*

1.2.2 Examples

A binary-state testing problem

A sender (a product designer) wishes to prove the quality of an innovation by submitting test data to a receiver (the CEO). The innovation may improve on an existing product or not: its contribution to the profitability of the product is either high or low ($\Theta = \{\theta_H, \theta_L\}$) with $\theta_H = 1$ and $\theta_L = 0$. The sender persuades the receiver that the quality is high by submitting results from a limited trial where the outcomes are $1 =$ *Don't click ad*, $2 =$ *Click ad but don't purchase*, and $3 =$ *Click ad and purchase*. It is commonly known that in the two states of the world, the true likelihood of each outcome is as follows:

j	$f_j(1)$	$f_j(2)$	$f_j(3)$
θ_H	1/2	1/5	3/10
θ_L	13/20	3/20	1/5

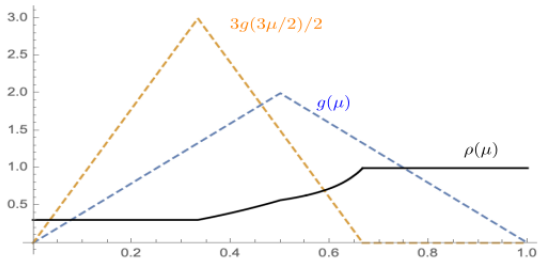
Table 1.1: The generating distribution of outcomes under states θ_H and θ_L.

Imitation implies that every on-path message either contains data distributed like f_H or like f_L, which the receiver can interpret as a claim that "the state is θ_H" or "the state is θ_L", respectively. But since the sender strictly prefers the receiver to believe the state is θ_H with higher probability, there is no reason to imitate f_L. Indeed, part 1.2.1(b) of the definition of an imitation equilibrium ensures that the only on-path messages take the form μf_H, since no posterior belief of the receiver is worse for the sender than full certainty that $\theta = \theta_L$.

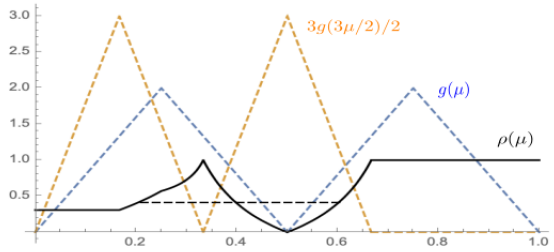
Additionally, the sender chooses an amount of data to send, which the receiver can interpret as an amount of support to back up their claim. The sender's true dataset determines whether they are able to submit more or less data that fits the distribution, and it is optimal for the receiver distinguish them along this margin to encourage partial separation. When evidence is generated as in Table 1.1, the sender can send $m = \mu f_H$ if and only if the true data are $\mu' f_H$ with $\mu' \geq \mu$, or $\mu' f_L$ with $\mu' \geq \frac{3}{2}\mu$. The *distinguishability factor* of $\frac{3}{2}$ reflects the relative advantage to a sender under θ_H of imitating f_H, and comes from the fact that in order to be able to submit enough observations of outcome 3 to imitate μf_H, a sender under θ_L must start with $\frac{3}{2}$ as much data.

As a naive first guess, suppose that the sender's strategy is to always send the maximum possible amount of data that is distributed like f_H.

$$m_{max}(\mu f_H) = \mu f_H, \quad m_{max}(\mu f_L) = \frac{2}{3}\mu f_L. \tag{1.4}$$



(a) Posterior beliefs when g is triangular



(b) Posterior beliefs when g is double-triangular

Figure 1.2: Inferences from message $m = \mu f_H$ in the binary-state example.

Consider the uniform prior $\beta_0(\theta_H) = \frac{1}{2}$ and a data-mass distribution that is "triangular",

$$g(\mu) = 2 - 4|x - 1/2|.$$

The receiver's inference upon receiving a message $m_{max} = \mu f_H$, plotted by the solid line in Figure 1.2(a), is

$$\rho_{\max}(\mu) = \begin{cases} 1, & \mu \geq 2/3 \\ \frac{4-4\mu}{10-13\mu}, & \mu \in [1/2, 2/3) \\ \frac{4\mu}{6-5\mu}, & \mu \in [1/3, 1/2) \\ \frac{4}{13}, & \mu < 1/3. \end{cases}$$

To visualize how the receiver constructs the posterior inference, observe that the density of senders who send a message μf_H for a μ for whom the true state is θ_H and θ_M are $g(\mu)$ and $\frac{3g(3\mu/2)}{2}$, which are plotted as two dotted lines. Their ratio is the likelihood ratio of the high vs. the low state given message μf_H.

Observation 1. *$\rho_{\max}$ depends only on β_0, g, and the distinguishability factor.*

In other words, the distinguishability of f_H from f_L is a sufficient statistic for both distributions that captures their implications for inferences under the naive strategy. In fact, we can verify that the naive messaging strategy in eq. 1.4 supports an equilibrium, under the assumption that any off-path messages feasible for some low-state type of the sender are evidence of the low state. More generally, the naive strategy is the unique imitation equilibrium strategy whenever it induces monotone inferences from the receiver.

In some cases, $\rho_{max}(\mu)$ is nonmonotone, such as when μ takes the "double triangular" distribution

$$g(\mu) = \begin{cases} 2 - 8|x - 1/4|, & x \in [0, 1/2] \\ 2 - 8|x - 3/4|, & x \in (1/2, 1]. \end{cases}$$

If all types of the sender send the maximal mass of data imitating f_H, then the message $1/2 f_H$ makes the receiver more pessimistic than the message $1/3 f_H$, and incentive compatibility fails because a sender who was to send the former would choose to send the latter instead.

This is easily fixed, however, if, within a pooling interval, all types of the sender still imitate f_H, but send less than the maximal mass. The dashed line in Figure 1.2(b) shows that the receiver's inferences given all messages in an interval can be equalized this way, so that the unique equilibrium inference is instead an ironed version of $\rho_{\max}(\mu)$.[8]

Remark. *In this example, data on outcome* 3 *is the limiting factor that restricts the state-θ_L sender from providing more data to support "the state is θ_H". Data on the remaining outcomes does not matter. There are several things one could do with the remaining outcomes that would not change the results of disclosure:*

1. *Merge them, i.e., design an experiment that does not distinguish between outcomes* 1 *and* 2*, and let the sender self-disclose raw data generated from the simplified experiment.*
2. *Delete them, i.e. allow the sender to report only observations of outcome* 3*, while leaving no option to input instances of outcomes* 1 *and* 2*.*

Finding the determinants of distinguishability therefore points out ways to lighten the burden of data transmitted while retaining the same information under cherrypicking.

A 3-state extension

Now suppose there is a 3rd possible state for the innovation that corresponds to adding a new feature that makes consumers very willing to explore the product, but not much more likely to purchase it (i.e., the product goes viral). It is represented by the state θ_M. The medium-quality innovation yields a different joint distribution of views and purchases; to summarize, the distributions of the same 3 outcomes under all states are given by Table 1.2.

j	$f_j(1)$	$f_j(2)$	$f_j(3)$
H	1/2	1/5	3/10
M	3/10	19/40	9/40
L	13/20	3/20	1/5

Table 1.2: Data-generating distributions under states θ_H, θ_M and θ_L

[8]The ironing process can be described as follows. If all types that would send $m = \mu f_H$ for some $\mu \in [\underline{\mu}, \bar{\mu}]$ were pooled, the receiver's inference given the pool would be

$$p(\underline{\mu}, \bar{\mu}) = \frac{\int_{\underline{\mu}}^{\bar{\mu}} g(\mu) d\mu}{\int_{\underline{\mu}}^{\bar{\mu}} \left(g(\mu) + \frac{3g(3\mu/2)}{2} \right) d\mu}.$$

Given some μ^* at which $\rho_{\max}(\mu)$ is decreasing, we can find $\underline{\mu} < \mu^* < \bar{\mu}$, such that either $\rho_{\max}(\mu)$ is increasing at both $\underline{\mu}$ and $\bar{\mu}$, and $\rho(\underline{\mu}) = \rho(\bar{\mu}) = p(\underline{\mu}, \bar{\mu})$; or $\underline{\mu} = 0$ and $\rho_{\max}(\mu)$ is increasing at $\bar{\mu}$ with $\rho(\bar{\mu}) = p(\underline{\mu}, \bar{\mu})$; or, $\bar{\mu} = 0$ and $\rho_{\max}(\mu)$ is increasing at $\underline{\mu}$ with $\rho(\underline{\mu}) = p(\underline{\mu}, \bar{\mu})$. There is a pair $(\underline{\mu}, \bar{\mu})$ satisfying these criteria that are closest to μ^*, and they are the endpoints of the ironing interval.

Consider first the problem of a sender who knows that $\theta = \theta_L$. There are now 2 distributions that they can imitate: f_M and f_H. On the other hand, a sender for whom $\theta = \theta_M$ may wish to imitate is f_H, but never f_L. It takes at least $\frac{5}{4}\mu f_L$ and $\frac{19}{6}\mu f_L$ to imitate μf_M and μf_H, respectively, and $\frac{5}{3}\mu f_M$ to imitate μf_H. We can now keep track of three distinguishability factors, $r_L(M) = \frac{19}{6}$, $r_L(H) = \frac{3}{2}$, and $r_M(H) = \frac{5}{3}$.

Relative to the binary-state case, solving for the equilibrium when $|\Theta| \geq 3$ involves an extra step: understanding which state a sender will choose to target in imitation. Nevertheless, construction can proceed from the top down. First observe that types μf_H with $\mu > \frac{1}{r_L(H)}$ can separate and obtain a payoff of θ_H. We then ask which types of senders obtain a payoff $v \in (\theta_M, \theta_H)$. For this restricted set of payoff frontiers, it suffices to consider imitating f_H only, since no message imitating f_M can yield a payoff greater than θ_M. Similarly to the binary-state case, in this regime the receiver can conjecture that the sender "imitates as much of f_H as possible", and restore monotonicity if needed by ironing. For payoff frontiers corresponding to $v < \theta_M$, one of two things is possible. If the state is θ_M and the sender has enough data to separate from all other types that cannot obtain $v > \theta_M$ by imitating f_H, then they play their full dataset and separate. Otherwise, unless the state is θ_L, the sender plays their full dataset, but their full dataset is imitated by some type for whom the state is low, and who plays a strategy that mixes between imitating f_H and f_L. Figure 1.3 summarizes how the three distinguishability factors $r_L(M), r_L(H)$, and $r_M(H)$ determine the equilibrium: it projects all types onto a space that summarizes how imitable f_H and f_M are, as the vertical and horizontal dimensions, and shows their imitation strategies and payoffs in equilibrium. Note that distinguishability between θ_H and the other two states is supported by outcome 3, while distinguishability between θ_M and θ_L is supported by outcome 2. An immediate consequence is that data on both outcomes are necessary to support the receiver-optimal equilibrium. The multiplicity of disclosed outcomes can lead to other, novel features.

Observation 2. *When $|\Theta| \geq 3$, sending observations of multiple outcomes may be necessary to support the receiver-optimal equilibrium. In addition, the equilibrium can have additional features that are not possible when $|\Theta| = 2$ or $|\mathcal{D}| = 2$:*

- *An interior mass of observations of some outcomes may be strictly optimal.*
- *Keeping μ constant, the sender can receive greater payoffs under a lower state.*

As an example of the first point, suppose that $u_{\sigma^*}(\mu f_H)$ is strictly increasing in μ. For μ close to 1, consider type μf_L imitating $\frac{2}{3}\mu f_H$ by sending a mass $\frac{1}{3}\mu$ of observations of outcome 1. Sending a greater mass would rule out the type $\frac{2}{3}\mu f_H$ that it wants to imitate, but sending less would rule in types like $(\frac{10}{9} - \epsilon)\mu f_M$, which would worsen the receiver's inference from the message. For the second point, observe that the type μf_L obtains a greater payoff than the type μf_M when $\mu = 1$: the former can imitate $\frac{2}{3}\mu f_H$, while the latter can only imitate $\frac{1}{2}\mu f_H$. In this case, the true state determines a sender's welfare not directly through the receiver's best response to it, but through the relative advantage it confers in matching *better* states on observables.

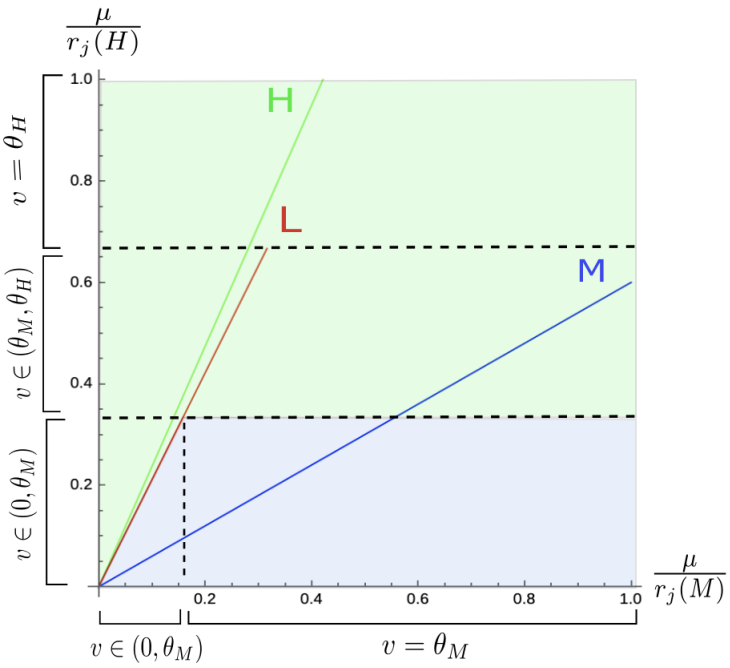


Figure 1.3: Equilibrium imitation strategy under the data-generating distribution in Table 1.2

A final novelty of the pooling equilibrium with 3, and indeed more, states is that the sender will separate and fully inform the receiver of the state only if they possess an intermediate amount of data – ignoring the best and worst state, under θ_M there is a temptation to drop evidence with too much data, and an inability to distinguish oneself from imitators when too little data is acquired.

1.3 Construction and characterization

This section characterizes the imitation equilibrium, constructs it, and shows that it is essentially unique. The imitation equilibrium is distinguished among equilibria by the fact that in it, worse types imitate better types (condition 1.2.1b). This is directly reflected in the structure of the receiver's beliefs once they receive an on-path message m: the best case for any message is that the receiver takes it literally to be the sender's full dataset, while any skepticism that this is true negatively affects their inferences. Any off-path dataset $m \in \mathcal{T}$ might as well be taken literally,

$$q^*(\cdot|m) = \mathbb{1}_m \text{ for all off path } m \in \mathcal{T}, \tag{1.5}$$

and is off path not because the receiver's inferences are "artificially depressed" but because imitating some other dataset is strictly preferred for the type $t = m$. Therefore, the sender benefits from selective disclosure if and only if they lie – there are no imitation equilibria

that increase the payoff of truthful senders relative to their payoff when the receiver is fully informed. On the other hand, truthful senders can suffer – since other senders can dishonestly imitate them, the receiver can be skeptical of their dataset even if they tell the truth.

In addition, for any dataset *not* resembling some raw dataset, $m \notin \mathcal{T}$, there are off-path beliefs

$$q^*(\cdot|m) = q^*(t| \operatorname*{arg\,min}_{t' \supseteq m,\ t' \in \mathcal{T}} \mathbb{E}_{\beta(\cdot|\sigma^*(t'))}[\theta]) \text{ for all } m \in \mathcal{M} \setminus \mathcal{T}, \tag{1.6}$$

and given these beliefs, senders never benefit from playing a dataset that the receiver knows for sure to be incomplete. Because of this, an observer of the interaction between senders and receivers would not be able to tell if senders are strategically omitting data simply by looking at the distributions of the published data – some prior about how much data the sender ought to have is necessary to know if observations are being dropped.

We have established that, in an imitation equilibrium, a sender's ability to positively influence the receiver depends on the extent to which they can imitate another state. In turn, this depends on the mass of their own dataset, μ, and the extent to which f_k can be distinguished from f_j, which is given by

$$r_j(k) = \max_{d \in \mathcal{D}} \frac{f_k(d)}{f_j(d)}.$$

This distinguishability factor $r_j(k)$ is a measure of the comparative advantage to a sender under state θ_k to reporting a dataset distributed like f_k, relative to a sender under state θ_j.[9] It can be interpreted to mean that "under state θ_j, a sender would need $r_j(k)$ times as much data to imitate μf_k than under θ_k". A sharp feature of the continuum model is that pairwise distinguishability comparisons fully suffice to summarize the impact of the shape of generating distributions $\{f_j\}_{j=1}^J$ on the imitation equilibrium outcome.

Proposition 1 (Sufficiency)**.** *Two games $\mathcal{G}$ and $\mathcal{G}'$ must yield the same outcome if they share the same state space Θ and priors β_0 and G, and for all j and k,*

$$\max_{d \in \mathcal{D}} \frac{f_k(d)}{f_j(d)} \equiv r_j(k) = r'_j(k) \equiv \max_{d' \in \mathcal{D}'} \frac{f'_k(d')}{f'_j(d')}.$$

In other words, even if $\mathcal{D}$ is very large, $\{f_j\}_{j=1}^J$ only affect the menu of possible beneficial manipulations through a select set of summary statistics, which are each supported by a single point in $\mathcal{D}$. We will delay discussion of the comparative statics of distinguishability, as well as their implications for optimal experimental design, to section 1.5. In the present section, we leverage these factors to complete our characterization of the imitation equilibrium. All equilibrium outcomes can be described by a vector-valued function $\hat{\mathbf{u}}(\mu) = (\hat{u}_j(\mu))_{j=1}^J$, with $\hat{\mu}_j(\mu) = u_\sigma(\mu f_j)$. The imitation equilibrium outcome has an even simpler description: each

[9] Equivalently, we can consider its inverse, $\frac{1}{r_j(k)}$, an *imitability* factor that describes how easily f_k is imitated under state θ_j.

sender's messaging problem can be simplified down to the choice of a weakly better state to imitate, $k \in \{j, \dots, J\}$, and an amount μ of that state's distribution to send – "as much as possible" is always weakly optimal, though, as with ironing in example 1.2.2, may not be the only strategy played in equilibrium. Since $\hat{u}$ describes the payoff under every state to every μ, its inverse $\hat{\boldsymbol{\mu}}$, defined as

$$\hat{\mu}_j(u) = \min\{\mu : \hat{u}_j(\mu) \geq u\},$$

describes a *burden of proof* in order to achieve payoff u, and what is necessary is that a type t can provide at least a measure $\hat{\mu}_k(u)$ of distribution f_k, where $\theta_k \geq u$. Crucially, fixing the pairwise distinguishability factors, optimality of the sender's imitation strategy amounts to saying that a sender that achieves payoff u via imitation is either truthful with $\theta_j \geq u$ or imitates another state $\theta_k > u$ in the set

$$A_j(\mu) = \left\{ \theta_k : k \in \arg\max_{k>j} \hat{u}_k \left(\frac{\mu}{r_j(k)} \right) \right\}.$$

Theorem 2 (Existence and uniqueness)**.** *There there exists an essentially*[10] *unique imitation equilibrium, implemented by a vector-valued burden of proof function* $\hat{\boldsymbol{\mu}} : [0, \theta_J] \to \mathbb{R}^J$ *with outcome* $\hat{\mathbf{u}}$ *such that*

1. $\hat{u}_j(\mu)$ *is continuous and (weakly) increasing in* μ *for all* j*.*
2. $\sigma^*(\mu f_j)$ *is supported on* $\left\{ \mu' f_k : \mu' = \hat{\mu}_k \left(\hat{u}_k \left(\frac{\mu}{r_j(k)} \right) \right) \text{ and } \theta_k \in A_j(\mu) \right\}$.

1.3.1 Construction of the equilibrium

In the Appendix, we give the details of the step-by-step construction of σ^* in general. But to capture the main idea, consider a minimal setup that illustrates the forces at play. Suppose we face the problem of constructing $\hat{\boldsymbol{\mu}}(v)$ for $v \in [\theta_{J-1}, \theta_J]$, assuming that $\hat{\boldsymbol{\mu}}(\theta_{J-1})$ is known. Fig. 1.4 shows that a typical type space can be projected onto 2 dimensions: one dimension describes the ability of each type to imitate f_J, given by $\frac{\mu}{r_j(J)}$, and the other dimension describes their ability to imitate f_{J-1}, given by $\frac{\mu}{r_j(J-1)}$. We can plot senders with all possible amounts of data under a given state as a ray when we describe the type space this way. Since any $v > \theta_{J-1}$ is obtained through imitating one of these two types, this description is sufficient to determine the imitation strategies used to obtain this subset of responses from the receiver.

The burden-of-proof vector lies in the same space and describes two simple things: which of the two states each type imitates, and what the highest action is that they can induce the receiver to take by doing so. A couple of observations allow us to identify the unique continuation of $\hat{\boldsymbol{\mu}}(v)$ at and to the left of any v^* whenever $\hat{\boldsymbol{\mu}}(v)$ is already known for all $v > v^*$.

[10] β^* is uniquely determined, and σ^* is uniquely determined up to payoff-irrelevant mixing probabilities.

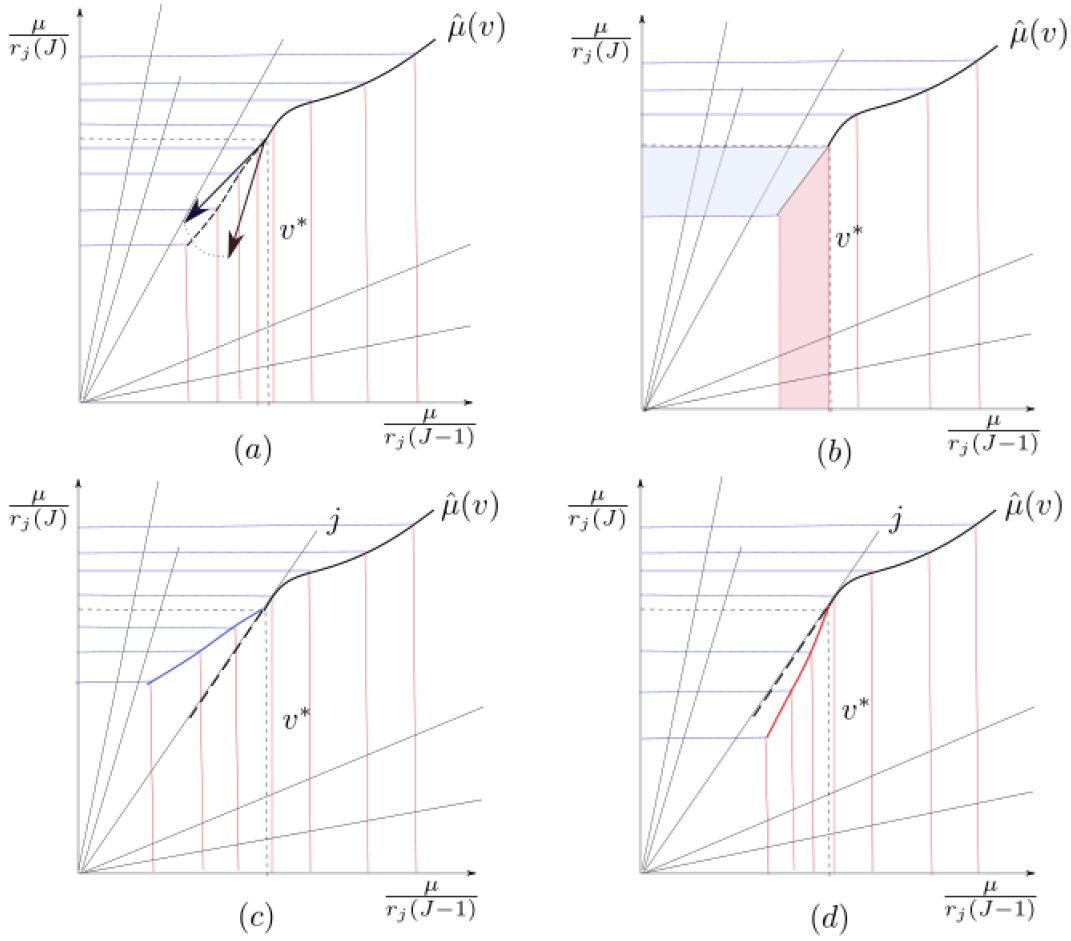


Figure 1.4: Example of $\hat{\boldsymbol{\mu}}(v)$ equalizing payoffs to imitating each state.

Taking the higher payoff frontiers to be fixed, focus on the set of types unable to meet any component of $\hat{\boldsymbol{\mu}}(v)$ for any $v > v^*$. There may exist within this set a self-separating set of positive measure that can pool with each other to induce action v^*. Fig. 1.4(b) shows that if so, $\hat{\boldsymbol{\mu}}(v)$ is discontinuous at v^*, since the equilibrium construction then immediately pools these types and assigns them all a payoff of v^*. Otherwise, $\hat{\boldsymbol{\mu}}(v^*) = \lim_{\epsilon \to 0} \hat{\boldsymbol{\mu}}(v^* + \epsilon)$.

The key fact is that given $\hat{\boldsymbol{\mu}}(v^*)$, it is always possible to exactly specify $\hat{\boldsymbol{\mu}}(v)$ for v in some, possibly small, nonempty interval $(v^* - \Delta, v^*)$. Consider first the case in which all types in the v^*-payoff frontier strictly prefer to imitate either f_J or f_{J-1}. When $\hat{\mu}_j(v^*)f_j$ imitates distribution f_J, then all types μf_j with μ close to $\hat{\mu}_j(v^*)$ behave likewise, and the same is true for those imitating distribution f_{J-1}. In other words, the payoff frontiers are locally determined because imitation strategies are fixed, up the amount of data submitted. Panel (a) of Fig. 1.4 shows that $\hat{\boldsymbol{\mu}}(v)$ then follows along the path of equivalent payoffs from imitating either state, and is continuous, due to the continuity of g.

A second possibility is that for some j, type $\hat{\mu}_j(v^*)f_j$ may indeed be indifferent between imitating f_J and f_{J-1}, and mixes between the two with interior probability. Locally, for μ close to $\mu_j(v^*)$, the types μf_j must also be indifferent, and so for a set of values $v \approx v^*$, $\hat{\boldsymbol{\mu}}(v)$ coincides with the set of types under state θ_j that achieve the corresponding payoff. For all state-θ_j senders obtaining a payoff in this range, the mixed strategy played equalizes payoffs to imitating each of the two highest states. Fig. 1.4(c) shows that if $\sigma(\frac{\mu}{r_j(J)} f_J | \mu f_j)$ increases too quickly, this fails to hold, since then payoffs to imitating θ_J decrease quickly relative to those to imitating θ_{J-1}, and (d) shows that payoffs to imitating θ_J decrease too quickly in the opposite case. There is, then, a unique continuation of the mixed strategy that respects the restriction on $\hat{\boldsymbol{\mu}}$, and it is continuous due to the continuity of g.

When there are more than 2 candidate states to imitate, the construction is slightly more complicated in that there may be more than one state under which types are indifferent across distributions to imitate, and a given type may be indifferent between imitating more than 2 different states. Nevertheless, the idea is the same. It is always possible to construct an interval of frontiers and their associated equilibrium strategies, given knowledge of higher-payoff frontiers. The construction technique then proceeds interval-by-interval, where we note that each interval formed in a step of the process is nonempty but may be small: it may be necessary to switch from handling the problem as in the first case to handling it as in the second case, and vice versa, multiple times as the algorithm proceeds to successively lower payoff frontiers.

1.3.2 A separation theorem

Let us return briefly to the matter of why the imitation outcome stands out from other equilibrium outcomes. It turns out that, although we can construct the imitation equilibrium payoff frontiers iteratively, we can also characterize them each individually, and independently of the remainder of the equilibrium. Put simply, imitation equilibrium payoff frontiers universally divide the type space into a greater-value upper region and a lesser-value

lower region, and they are the only frontiers to do so.

We start with some definitions.

Definition 1.3.1. *An* upper pool *of payoff frontier* $\hat{\boldsymbol{\mu}}(v)$ *is a set*

$$\bar{T} = U(\hat{\boldsymbol{\mu}}(v)) \setminus U(M)$$

for some collection of messages M*.*

Definition 1.3.2. *A* lower pool *of of payoff frontier* $\hat{\boldsymbol{\mu}}(v)$ *is a set*

$$\underline{T} = U(M) \setminus U(\hat{\boldsymbol{\mu}}(v))$$

for some collection of messages M*.*

An upper pool consists of all types above the payoff frontier $\hat{\boldsymbol{\mu}}(v)$ but below some other frontier, while a lower pool consists of types below it but above another frontier.

We define the pooled value of any set of types, $u_{pool}(T)$, to be the receiver's expectation of the state given that the sender's type is in the set T, and state the separation theorem:

Theorem 3 (Separation)**.** *For any nonempty upper pool* $\bar{T}$ *and lower pool* $\underline{T}$ *of* $\hat{\boldsymbol{\mu}}(v)$*,*

$$u_{pool}(\bar{T}) \geq v > u_{pool}(\underline{T}).$$

In other words, upper pools are weakly improving and lower pools are strictly worsening — for any subset of $\mathcal{T}$ that is bounded by two frontiers and contains $\hat{\boldsymbol{\mu}}(v)$, the value of the part above $\hat{\boldsymbol{\mu}}(v)$ is at least v, while the value of the part below is less than v.[11]

The fact that upper pools are improving is a consequence of the conditions of imitation equilibria: the property holds because in each group of senders who send the same message under σ^*, only those with worse-than-average values can be truncated by excluding $U(M)$. On the other hand, the equilibrium we construct has worsening lower pools because in it, any potentially self-separating pool of senders below $\lim_{\epsilon \to 0} \hat{\boldsymbol{\mu}}(v + \epsilon)$ that achieves a value of at least v must lie above the frontier $\hat{\boldsymbol{\mu}}(v)$.

These properties guarantee uniqueness of the imitation equilibrium outcome if we use them to compare outcomes under σ^* and another PBE, σ. If the outcome under σ differs from that under σ^*, then worsening lower pools under σ^* imply that there is a frontier with a worsening upper pool under σ. Moreover, the only frontiers in $\mathcal{T}$ that satisfy either property are the frontiers of σ^*. Given any prospective frontier and its associated payoff, checking either of these properties in isolation is enough to verify that it shows up in the imitation equilibrium, and may in some cases be easier than constructing the entire imitation equilibrium outcome.

[11] The former inequality is weak and the latter strict because we have defined the imitation payoff frontiers such that, when there are multiple types μf_k that all achieve v, $\hat{\mu}_k(v)$ is the lowest such μ.

The separation theorem is a general result — it also applies to finite evidence games, where it is related to the "downward biased" characterization of Rappoport (2022). In all these cases, worsening lower pools rules out credible inclusive announcements, and improving upper pools turns out to imply that no other equilibrium is credible inclusive announcement-proof.

1.4 Comparative statics

All imitation outcomes share some concrete features. Here, we present comparative statics of the sender's reports in μ, of the sender's welfare with respect to the receiver's prior belief about θ and μ, and of separation as $Var[\mu] \to 0$. We begin with a corollary to Theorem 2.

Corollary (to Thm. 2)**.** *Under σ^*, there are thresholds $z_j^* > z_j^{**}$ 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 $\hat{u}_j(\mu) > \theta_j$.*
- *Whenever $\mu \in (z_j^{**}, z_j^*]$, the sender is honest and the receiver knows it upon receiving the data: $\hat{u}_j(\mu) = \theta_j$.*
- *Whenever $\mu \leq z_j^{**}$, the sender is honest, but the receiver believes they are a worse type with positive probability, and $\hat{u}_j(\mu) < \theta_j$.*

We can think of senders with $\mu > z_j^*$ as high-data senders, with enough data to benefit from manipulating their data against the receiver's uncertainty about their data endowment. The costs of voluntary disclosure are borne by low-data senders, those with $\mu < z_j^{**}$, who the receiver is skeptical of even when they are truthful. These thresholds vary by j, and in particular, $z_1^* = 0$ and $z_J^* = 1$. However, they need not be monotone in j.

The potential presence of an intermediate, full-information interval between disjoint upper and lower partial-pooling intervals when we fix θ and vary μ is a novel feature of these equilibria that occurs when there are multiple imitated states with different distinguishing outcomes. It is a consequence of the fact that it requires a strictly greater amount of data to benefit from imitating a different state than it does to send one's full dataset and discourage all imitators. The structure of pooling and separation contrasts with strategies in binary-state models of voluntary disclosure, or in models with ordered outcomes. In those cases, full separation only occurs at the very top, that is, for types with a maximal state and a maximal amount of evidence (see, for example, Dye (1985) and Dzuida (2011)). We show that this doesn't have to be true in general: although they remain able to separate, types with the most evidence are often more tempted to pool with others.

Partial pooling occurs in the receiver-optimal equilibria in our model because uncertainty about μ makes it impossible for lower-data, higher-state senders to separate from higher-data, lower-state senders. In the absence of uncertainty about μ (that is, if μ is commonly known to the sender and the receiver) the disclosure game is a case of the games studied

by Grossman (1981) and Milgrom (1981), in which unraveling occurs. The distribution of μ in our model, while assumed to be nondegenerate, can be arbitrarily close to a point mass, and outcomes converge to the full-information outcome as the receiver's uncertainty about μ vanishes.

Claim 2. *As $Var[\mu] \to 0$, we have $Pr(|\hat{u}_j(\mu) - \theta_j| > \epsilon) \to 0$ for all j.*

This should be unsurprising: when the receiver knows μ quite well, any dataset with fewer-than-expected observations is quite suspicious and is heavily discounted, and this limits the returns to omitting data.

1.4.1 Complementarity with public information

Next, we show that public information can complement voluntary disclosure. As a preliminary, observe that the receiver is worse off given uncertainty about either the payoff-relevant state, or its relation to the experiment. Suppose that there are any two games $\mathcal{G}(\Theta, \mathcal{D}, \beta_0, \{f_j\}_{j=1}^J, G)$ and $\mathcal{G}'(\Theta', \mathcal{D}, \beta_0', \{f_{j'}'\}_{j'=1}^{J}, G)$ that have the same outcome space $\mathcal{D}$ and data-mass distribution G, [12] but describe different sets of possible states. If the receiver knows that the true state might be in Θ or Θ', but is uncertain of which and has prior α, $1-\alpha$, respectively, of the likelihood of each case, then they can be modeled as playing a third game, $\mathcal{G}^{uc}$, in which the set of states is $\Theta^{uc} = \Theta \bigcup \Theta'$ with each state retaining its data-generating distribution. Their prior over Θ^{uc} is given by

$$\beta_0^{uc}(\theta_j) = \alpha\beta_0(\theta_j) \quad \text{for } \theta_j \in \Theta, \qquad \beta_0^{uc}(\theta_{j'}') = (1-\alpha)\beta_0(\theta_{j'}') \quad \text{for } \theta_{j'}' \in \Theta'.$$

Claim 3. *The receiver's expected payoff in $\mathcal{G}^{uc}$ conditional on $\theta \in \Theta$ is less than their expected payoff in $\mathcal{G}$, and strictly so if there is a type of the sender with state in Θ for which the outcomes differ in the two games.*

Greater ex-ante uncertainty, in any of a number of dimensions, is generally worse for the receiver. When the receiver suffers from not knowing whether the state is in Θ or Θ', it is because a state $\theta_{j'} \in \Theta'$ differs from a state $\theta_j \in \Theta$ for one or more of the following reasons:

1. Its numerical value is different;
2. The prior likelihood that the receiver assigns to it is different, $\beta_0(\theta_j) \neq \beta_0'(\theta_{j'})$;
3. The distribution of data it generates is different, $f_j \neq f_{j'}'$.

Applying Claim 3 to the first case shows that the receiver is worse off when they understand the set of possible states of the world corresponding to different distributions of experimental outcomes, but are uncertain about the optimal action to take even with full

[12] We prove the claim that follows in a more general setup, in which the the games need not have the same data-mass distribution and the distribution of μ can be state-contingent, that is, the two games have state-contingent data mass distributions $\{G_j\}_{j=1}^J$, $G_{j'}'\}_{j=1}^{J'}$, respectively.

knowledge of the distribution of the data. In the second case, we see that having an incorrect prior about the state of the world cannot benefit the receiver in expectation.

The third case is perhaps the most interesting: it captures a receiver's uncertainty about the distribution of outcomes conditional on the true, payoff-relevant state. We illustrate how this could play out in the context of our previous example (Ex. 1.2.2), by expanding the state space to include not only $\Theta = \{\theta_H, \theta_L\}$, but also $\Theta' = \{\theta_{H'}, \theta_{L'}\}$, where θ'_H and θ'_L are analogous to θ_H and θ_L, respectively, except that they represent data generated from a different, high-demand test market in which the rate of both views and sales is elevated relative to the baseline. Suppose that this difference in the test market does not affect the receiver's optimal action, so $\theta_{H'} = 1$ and $\theta_{L'} = 0$.

In addition, we consider randomizing impressions 50/50 to a control arm to the experiment, in which viewers are shown an original, unmodified version of the product instead. This replicates the format of an A/B test, and can allow the receiver to identify the the test market's demand conditions by observing how the original version of the product would have performed. Table 1.3 shows a scenario in which, the absence of controls, state θ_H and $\theta_{L'}$ are not distinguishable, even given a full dataset, since they generate identical data. However, with controls, θ is once again identified.

		Control			Treatment		
	j	$f_j(1)$	$f_j(2)$	$f_j(3)$	$f_j(4)$	$f_j(5)$	$f_j(6)$
Low demand	H	0.325	0.075	0.10	0.25	0.10	0.15
	L	0.325	0.075	0.10	0.325	0.075	0.10
High demand	H'	0.25	0.10	0.15	0.1625	0.1375	0.20
	L'	0.25	0.10	0.15	0.25	0.10	0.15

Table 1.3: Generating distribution of data in an A/B testing example with an uncertain test market.

When the sender can cherrypick the data, adding the control does not fully mitigate the effect of uncertainty between Θ and Θ'. It is almost immediate to see, for example, that if the receiver places 50% probability each on the event sthat the state is in Θ and Θ', then there is a smaller range of messages from which they can verify that the state is θ_H. If the sender sends $m = \mu f_H$ for $\mu \in (2/3, 10/13)$, a sender with mass $\mu' \in (8/9, 1)$ of data distributed like $f_{L'}$ could also have sent such a message, even though there is nobody type could have done so under state θ_L. Indeed, in this case, the receiver's uncertainty about the test market makes them strictly worse off.

Put differently, the receiver would always at least weakly benefit from observing a signal ϕ that tells them whether the game is $\mathcal{G}$ or $\mathcal{G}'$, and this is true of any finite public signal. A signal that is not informative about θ per se may still be valuable alongside voluntarily disclosed data if it informs the receiver about how to interpret the data, and the signal need not be fully informative about any aspect of the experiment to yield a strict benefit.

Claim 4. *Let ϕ be a finite-valued public signal that is only informative about $\mathcal{D}$ and $\{f_j\}_{j=1}^J$. It strictly benefits the receiver for some prior β_0 if and only if two distinct realizations $\hat{\phi}$ and $\hat{\phi}'$ induce games $\mathcal{G}$ and $\mathcal{G}'$ such that $\arg\max_d \frac{f_k(d)}{f_j(d)} \neq \arg\max_d \frac{f'_k(d)}{f'_j(d)}$ for some $k > j$.*

These kinds of signals complement disclosed data by making the receiver less susceptible to manipulations of their auxiliary beliefs through data omission. They might pertain to any jointly estimated covariates. In addition to resolving the baseline uncertainty captured in the example, public signals might also give information about the space of underlying outcomes, the likelihood of randomization to treatment or control groups, or the composition of the trial population – these types of information all improve the usefulness of voluntarily disclosed experimental data.

1.4.2 Impact of beliefs on the sender

When the receiver's belief about the ex-ante probability of a given state θ_j increases relative to others, the receiver's skepticism weakly increases for all messages that yield a higher payoff to the sender than full certainty of that state. The reverse is true of all messages that yield a lower payoff than θ_j. An increase in the probability of θ_j therefore "pulls" the receiver's action towards θ_j given any message, which has the consequence of decreasing ex-post payoffs for all types of the sender that would originally have achieved $\hat{u}_j(\mu) \geq \theta_j$, and increasing them if originally, $\hat{u}_j(\mu) \leq \theta_j$.

To formalize this, let $\mathcal{G}$ be a disclosure game with prior β_0 about θ and $\mathcal{G}'$ be a game that is identical except for the prior β'_0 which differs from β_0, with $\beta'_0(\theta_j) > \beta_0(\theta_j)$ and $\frac{\beta_0(\theta_k)}{\beta_0(\theta_{k'})} = \frac{\beta'_0(\theta_k)}{\beta'_0(\theta_{k'})}$ for all other k, k'.

Claim 5. *Suppose that $\hat{\mathbf{u}}, \hat{\mathbf{u}}'$ are imitation equilibrium outcomes of $\mathcal{G}$ and $\mathcal{G}'$, respectively. Then $\hat{u}_{j'}(\mu) \geq \hat{u}_{j'}(\mu)$ whenever $\hat{u}_{j'}(\mu) \geq \theta_j$, and $\hat{u}_{j'}(\mu) \leq \hat{u}_{j'}(\mu)$ whenever $\hat{u}_{j'}(\mu) \leq \theta_j$.*

Lastly, we point out that MLRP shifts in the receiver's beliefs have a monotone impact on the sender's welfare. Simply put, every type of the sender benefits from a monotone likelihood ratio shift in the receiver's belief about the state, and suffers from a monotone likelihood ratio shift in their belief about the data-mass distribution. Intuitively, an upwards shift in the prior distribution in θ makes the receiver more willing to believe a claim that the state is high, and we show this formally in Appendix A.2. Such a shift unambiguously benefits the sender both conditional their realized dataset, and ex-ante. On the other hand, when the receiver expects μ to be greater, they are more *skeptical*: they infer a greater likelihood that a given message may have been selected from a larger dataset.[13]

Claim 6. *If two disclosure games $\mathcal{G}$ and $\mathcal{G}'$ are identical except for priors $\beta_0 \leq_{MLRP} \beta'_0$ and*

[13] Rappoport (2022)'s result can be used to show that the latter holds in finite-data games viewed as an instance of an abstract evidence game, and a similar argument shows that this is directly true in the continuum.

$g \geq_{MLRP} g'$, *then*

$$\hat{u}_j(\mu) \leq \hat{u}'_j(\mu) \quad \forall \mu f_j \in \mathcal{T}.$$

1.5 Experimental design

Our results highlight that the quality of the information the receiver obtains depends on how the data-generating process distinguishes states. This section focuses on interventions that aim to maximize distinguishability, and proposes a framework for optimally designing experiments to allow the receiver to extract payoff-relevant information from the sender through voluntary disclosure. In our model, an *experiment* is the data-generating process that provides the sender with their raw dataset, and is captured by a tuple $\mathcal{E} = (\mathcal{D}, \{f_j\}_{j=1}^J)$ consisting of the space of reported outcomes and the generating distribution of data over them. We assume that the remaining primitives of the game – state space, payoffs, and priors – are fixed, and consider the effect of varying the experiment that the sender observes.

A key fact is that whenever an experiment makes states pairwise more distinguishable, the receiver's welfare improves. Intuitively, increasing distinguishability allows higher-state types to separate themselves more effectively from lower-state types who would imitate them. The resulting equilibrium does not better separate every type from every other type – indeed there are types that would play different messages under one experiment that would play the same message in the other, in both directions – but, given the receiver's single-peaked expected utility, the more distinguishing experiment always makes the receiver better able to target the optimal action.[14]

Proposition 4 (Improvement)**.** *Suppose two experiments $\mathcal{E}$ and $\mathcal{E}'$ yield imitation equilibrium actions a and a', respectively.*

- *If $r'_j(k) \geq r_j(k)$ for all $k > j$, then $\mathbb{E}_{a,\theta}[u_r(a)] \leq \mathbb{E}_{a',\theta}[u_r(a')]$.*
- *If, in addition, $r'_j(k) > r_j(k)$ for some j, k such that there is some μf_j imitating $\hat{\mu} f_k$ under the imitation equilibrium with experiment $\mathcal{E}$, then $\mathbb{E}_{a,\theta}[u_r(a)] < \mathbb{E}_{a',\theta}[u_r(a')]$.*

In fact, by making a state arbitrarily distinguishable from others, we can guarantee that a sender under that state elicits at least their full-information action with high probability: the imitation equilibrium guarantees that . In the limit as all states become highly distinguishable, the receiver also approximately attains their full information payoff. On the other hand, if all states are negligibly distinguishable, the receiver learns essentially nothing, even as the sender is fully informed.

[14]The proof that distinguishability improves payoffs uses the fact that a mechanism designer that takes a sender's submitted dataset as a report is weakly more constrained by a sender's ability to deviate to sending a false dataset if the experiment has poor distinguishability. If the receiver does not have single-peaked preferences, then the imitation equilibrium outcome and the outcome of the optimal mechanism do not necessarily coincide, and increasing distinguishability may force the receiver to take a higher action after observing a message that few low-state types can imitate, when they would instead like to commit to responding to it with a lower action.

Claim 7. *With very high and very low distinguishability, outcomes approach those under full information and no information, respectively: for all $\epsilon, \delta > 0$,*

- *There exists $\underline{R} < \infty$ such that if $r_j(k) > \underline{R}$ for all $j < k$, then $Pr(|\hat{u}_k(\mu) - \theta_k| > \epsilon) < \delta$*
- *There exists $\bar{R} > 1$ such that if $r_j(k) < \bar{R}$ for all $j < k$, then $Pr(|\hat{u}_k(\mu) - \mathbb{E}_{\beta_0}[\theta]| > \epsilon) < \delta$*

where the likelihood is taken over realizations of μ.[15]

One way to better distinguish two states is to undertake a more detailed experiment. Without changing the experimental technology – that is, the underlying likelihood of events under different states – a researcher could investigate and record a more detailed set of outcomes in order to obtain finer data. To formalize this, suppose there is an existing outcome space $\mathcal{D}$, and consider a notion of a more elaborate outcome space $\mathcal{D}'$ that the researcher can obtain by splintering an existing outcome into multiple sub-outcomes to track.

Definition 1.5.1. *If there are two experiments $\mathcal{E} = (\mathcal{D}, \{f_j\}_{j=1}^J)$ and $\mathcal{E}' = (\mathcal{D}', \{f'_j\}_{j=1}^J)$ and a partition $\mathcal{P} = \{P_d\}_{d \in \mathcal{D}}$ of $\mathcal{D}'$ such that*

$$\sum_{d' \in P_d} f'_j(d') = f_j(d)$$

for all d in $\mathcal{D}$, then $\mathcal{E}'$ splinters *the outcome space of $\mathcal{E}$ and $\mathcal{E}$* merges *the outcome space of $\mathcal{E}'$.*

Immediately, we observe that for all θ_j and θ_k,

$$\max_{d' \in P_d} \frac{f'_k(d')}{f'_j(d')} \geq \frac{f_k(d)}{f_j(d)},$$

and so $r_j(k) \geq r'_j(k)$ whenever $\mathcal{D}'$ splinters $\mathcal{D}$.

Claim 8. *Splintering the outcome space weakly improves the receiver's expected payoff.*

In some cases, there is a most elaborate possible experiment $\mathcal{E}^*$, i.e., one that is a splintering of every other possible experiment. Suppose that costs and constraints on gathering, storing, and transmitting data are negligible. Then it is optimal for a designer who acts on behalf of the receiver to choose the most elaborate possible experiment. If instead the sender chooses the experiment, then the receiver should, if possible, incentivize the sender to choose the most detailed experiment by committing to accept nothing else. Since they follow a simple rule of thumb, these recommendations don't require detailed knowledge of the true data-generating process, and would be easy for even an uninformed designer to implement.

On the other hand, in practice there is often no binding limit to the number of ways that an experiment can be refined and complicated, at ever increasing cost. Despite the fact it never

[15] Despite the fact that $\bar{R} < \underline{R}$, we use this notation because $\bar{R}$ is an upper bound and $\underline{R}$ is a lower bound on the distinguishability sufficient for each case.

hurts, further splintering a dataset does not always strictly improve distinguishability. With precise information about the data-generating process, Proposition 1 allows us to identify instances when it is without loss to the receiver to merge outcomes relative to $\mathcal{E}^*$.

Proposition 5 (Merging). *Suppose that* $\mathcal{E}^* = (\mathcal{D}^*, \{f_j^*\}_{j=1}^J)$. *Let* $S^* = \bigcup_{j<k} \arg\max_d \frac{f_k^*(d)}{f_j^*(d)}$. *Then merging all outcomes in* $\mathcal{D}^* \setminus S^*$ *does not change the imitation equilibrium outcome.*

The set S^* consists of all outcomes that maximally distinguish one state from another. Mmerging other outcomes is without loss because S^* is sufficient to maximize every distinguishability factor in $\{r_j(k)\}_{j<k}$. We are left, generically, with a minimal experiment that suffices to reveal as much payoff-relevant information as possible to the receiver robustly over all possible priors.

Claim 9 (Minimality). *Fix* u_r, Θ, *and* g, *suppose that* $\mathcal{E}$ *is obtained from* $\mathcal{E}^*$ *by merging* S^* *and that* $\mathcal{E}'$ *merges some outcomes in* $\mathcal{E}$, *and suppose that* $\arg\max_d \frac{f_k^*(d)}{f_j^*(d)}$ *is unique for all* $j < k$.

Then there exists β_0 *such that the receiver is strictly better off with* $\mathcal{E}$ *than with* $\mathcal{E}'$.

This simplification of the experiment can be quite drastic, and in some familiar cases, including the case of a binary state space or an outcome space ordered by the monotone likelihood ratio property (MLRP), S^* is a singleton with only one "good news" outcome that maximally distinguishes higher states from lower ones, while all other outcomes in $\mathcal{D}^*$ can be merged and essentially ignored.[16] Formally, we say that $\mathcal{D}^*$ satisfies MLRP with respect to $\{f_j^*\}_{j=1}^J$ if, for any $j < k$ and $d < d'$,

$$\frac{f_k^*(d')}{f_j^*(d')} > \frac{f_k^*(d)}{f_j^*(d)}.$$

It is straightforward to see that MLRP implies that S^* comprises of the single maximal element in $\mathcal{D}^*$.

Even when $J > 2$, some degree of dimensionality reduction is often possible, especially if $J << |\mathcal{D}^*|$. In general, $|S^*| \leq \frac{J(J-1)}{2}$. The 3-state example 1.2.2 gives an instance in which this bound is tight because the maximizer, $\arg\max_d \frac{f_k(d)}{f_j(d)}$, is unique for all pairs $j < k$.

Corollary (to Prop. 5). *The minimal optimal experiment tracks at most* $\frac{J(J-1)}{2}+1$ *outcomes, and furthermore, if* $\mathcal{D}^*$ *satisfies MLRP with respect to* $\{f_j^*\}_{j=1}^J$, *then a binary outcome space suffices.*

[16] The imitation equilibrium in these cases has the same outcome as a sanitation equilibrium (Shin (2003)) in which the sender only reports observations of the outcome in S^*, and omits all others; however, it differs in that imitating senders generally report a positive mass of observations of these outcomes anyways, with no impact on the receiver's inferences.

1.6 Relationship to finite data

In the big picture, the purpose of modeling communication in this stylized, continuous-data disclosure game is to understand how senders will volunteer data in real-world disclosure settings, in which datasets are always finite. The comparative statics of section 1.4 and the experimental design implications of the previous section depend on the fact that datasets are well-described by μ and f_j, which is exactly true only in the continuum, but nearly true with large N in such a way that those results approximately carry over. This section makes precise the finite-data settings that we aim to approximate, and describes how the continuous-data model captures their regularities in the limit.

We model a sender who has access to a finite dataset of n i.i.d. observations drawn from $\mathcal{D}$ according to the state-contingent distribution f_j. The size of the sender's dataset is upper-bounded by N, but the sender may have access to $n < N$ observations as well, and the receiver is uninformed about how much data the sender has. Nature's sequence of moves in drawing the sender's dataset is: 1) draw the state, θ_j, according to prior β_0; 2) draw the number of observations, n, from distribution G_N; 3) for each of the n datapoints, draw their realized value i.i.d. from f_θ. Call the disclosure game with these parameters $\mathcal{G}_N(\Theta, \mathcal{D}, \beta_0, \{f_j\}_{j=1}^J, G)$. The data mass distributions $G_N(\cdot)$ capture the receiver's uncertainty about how much raw evidence the sender has, prior to selecting observations to reveal: for example, there may be uncertainty about the number of total trials in an experiment, or the number of trials out of N attempted that survived the entire trial period.

The sender's dataset is the empirical probability mass function $t = \frac{1}{N}(t_1, \ldots, t_D)$, where t_d is the number of observations of outcome d and $n(t) = \sum_{d=1}^D t_d$ is the number of observations they get. They are able to send any subset of their dataset as a message to the receiver, where

$$m \tilde{\subseteq} t \;\Leftrightarrow\; m_d \leq t_d \;\forall d \in \mathcal{D}.$$

In summary, the type space is $\mathcal{T}_N = \bigcup_{n=0}^N \mathcal{D}^n$, with type distribution

$$q_N(t) = \frac{n(t)!}{\Pi_{d=1}^D t_d!} \sum_{j'} \beta_0(\theta_{j'}) g_N(n(t)) \Pi_{d=1}^D f_{j'}(d)^{t_d},$$

and the message space $\mathcal{M}_N$ is identical to the space of types.

When datasets are finite, the sender's dataset does not perfectly inform them about the state: when f_j all have full support, any state is possible after observing any dataset. The likelihood of θ_j given that the raw dataset is t is

$$\pi_N(\theta_j | t) = \frac{\beta_0(\theta_j) g_N(n(t)) \Pi_{d=1}^D f_j(d)^{t_d}}{\sum_{j'} \beta_0(\theta_{j'}) g_N(n(t)) \Pi_{d=1}^D f_{j'}(d)^{t_d}},$$

and so, when the receiver observes a message and updates their belief about the sender's

type to $q_N(t|m)$, their posterior about the state updates to

$$\beta(\theta_j|m) = \frac{\sum_{t\in\mathcal{T}_N} q_N(t|m)\pi_N(\theta_j|t)}{\sum_{t\in\mathcal{T}_N} q_N(t|m)}. \tag{1.7}$$

We highlight that the distribution of datasets in the finite-data setting converges to the distribution of datasets in a continuous-data model. In particular, $g(\mu)$ represents the likelihood of obtaining a fraction μ of total potential data under state j, and analogously, $\frac{n}{N}$ is the fraction of total data available to the sender in the finite-data game. We can study a sequence of games such that as N increases, $NG_N(\frac{n}{N}) \to_{unif.} G(\mu)$, and note that if so, the type distributions also converge uniformly: $q_N \to_{unif.} q$.

Definition 1.6.1. $\mathcal{G}(\Theta, \mathcal{D}, \beta_0, \{f_j\}_{j=1}^J, G)$ *is the* limit game *for a sequence of finite-data games* $\{\mathcal{G}_N(\Theta, \mathcal{D}, \beta_0, \{f_j\}_{j=1}^J, G_N)\}_{N=1}^\infty$ *if* $NG_N(\frac{n}{N}) \to_{unif.} G(\mu)$.

Despite the fact that the type distributions converge, the type space $\mathcal{T}_N$ is drastically different from $\mathcal{T}$: in particular, $\mathcal{T}_N \sim \mathcal{M}_N$ and both approximately span a D-dimensional space of datasets for large N, while $\mathcal{T}$ is only 2-dimensional, as every dataset is described by μ and θ. While datasets far away from $\mathcal{T}$, that have distributions unlike the data-generating distribution in any state, become vanishingly unlikely as N grows large, they are never impossible except in the limit; this is why the continuum model is much easier to work with.

It remains possible to describe an imitation equilibrium and a truth-leaning equilibrium in the finite-data setting. The finite-data model is a special case of the evidence model in Hart et al. (2017) and Rappoport (2022). The former shows that truth-leaning equilibria exist and are unique and receiver-optimal in the finite-type setting, and also that they are always outcome-equivalent to imitation equilibria, although it does not guarantee that the strategies are equivalent. The latter includes an iterative algorithm to compute these equilibria; the number of steps is, however, exponential in $|\mathcal{T}_N|$, and as far as we can tell, there is no obvious way to obtain a significantly more efficient closed-form solution.

We can instead establish that the imitation equilibrium of the continuous-data model gives a perfect approximation to the limit outcome of communication in truth-leaning equilibria of finite-data games as $\mathcal{G}_N$ converge.[17] To make the comparison, the notion of an outcome should be extended across type spaces. There is a global data space $[0,1] \times \Delta\mathcal{D}$, invariant to N, that contains $\mathcal{T}_1, \ldots$ and $\mathcal{T}$ as long as they all share a space of observations. Recall that $u_{\sigma^*}(t)$, the outcome of the game for type t, is their payoff from the best feasible message given equilibrium beliefs. If $t \in [0,1] \times \Delta\mathcal{D}$, it need not also be in the literal type set for the outcome to be well-defined, since we can already infer whether t can feasibly send a

[17] We state the definition of convergence and the theorem below in terms of $N = 1, 2, \ldots$ rather than an arbitrary sequence of dataset sizes $N_1, N_2, \ldots$ only for the sake of notational brevity. The theorem applies just as well to any sequence of games $\{\mathcal{G}_{N_i}\}_{i=1}^\infty$ of increasing dataset size with uniformly convergent data distributions, since any such sequence is a subsequence of a convergent sequence of games $\{\mathcal{G}_N\}_{N=1}^\infty$.

message from the subset relation on $[0,1] \times \Delta\mathcal{D}$. The outcome to the hypothetical type can be understood as a thought experiment: "if the receiver believes we are playing a game with equilibrium σ^* or σ_N^*, and my dataset is t, what is the best payoff I can attain, even if t is inconsistent with the receiver's perceived game?"

Definition 1.6.2. *A sequence of equilibria* $(\sigma_1, \sigma_2, \ldots)$ *of games* $\{\mathcal{G}_N(\Theta, \mathcal{D}, \beta, \{f_j\}_{j=1}^J, G_N)\}_{N=1}^\infty$ *has outcomes that* converge *to the outcome of an equilibrium* σ *of the limit infinite-data game* $\mathcal{G}(\Theta, \mathcal{D}, \beta, \{f_j\}_{j=1}^J, G)$ *if the payoffs* $u_{\sigma_N^*}(t)$ *converge uniformly to* $u_{\sigma^*}(t)$ *over types in* $\mathcal{T}$.

Theorem 6. *If* $\mathcal{G}$ *is the limit game for finite-data games* $\mathcal{G}_1, \mathcal{G}_2, \ldots$ *with* $N = 1, 2, \ldots$ *respectively, then the truth-leaning equilibrium outcomes in* $\mathcal{G}_1, \mathcal{G}_2, \ldots$ *converge to the imitation equilibrium outcome of* $\mathcal{G}$.

Outcome convergence shows that it's reasonable to use the limit game to describe the distribution of actions the receiver takes after the sender discloses a large dataset, as well as the mapping from the truth to the receiver's inferences. At a high level, the proof follows from the convergence of type distributions $\mathcal{T}_N$ to $\mathcal{T}$, and from the separation theorem, which holds as well in truth-leaning equilibria of finite-data games. Appendix A.5 gives the formal argument and shows that the limit equivalence result partially extends to strategies, in addition to outcomes.

In addition, outcome convergence shows that previous sections' results on comparative statics and experimental design hold approximately for large finite datasets. When the number of observations is finite, splintering the data always leads to a strict improvement in the receiver's welfare, even when the outcome space already contains S^* and thus distinguishes the states as well as possible. However, in this case, the magnitude of the improvement vanishes and is negligible for large N. While merging non-distinguishing outcomes is only sharply optimal in the continuum, the convergence result guarantees us that it remains an actionable recommendation, yielding, in practice, nearly-optimal information to the receiver with minimally cumbersome datasets.

1.7 Conclusion

Inference under selective disclosure depends on an understanding of how data are generated, and how senders report – or omit – it. This paper underscores the simplicity of a receiver-optimal equilibrium reporting strategy for the sender: claim a possibly inflated state, and provide a large-enough body of evidence that supports it by mimicking the distribution of data it implies. Given their behavior, voluntary disclosure benefits precisely those senders who engage in strategic omission – those with a large amount of data or a low state – and worsens outcomes for those who are imitated – those who have less data, or a more desirable state to imitate.

Even when datasets are large enough to guarantee the sender is fully informed, strategic omission and uncertainty about exactly how much data the sender had mean that only

muddled information will reach the receiver. However, in the absence of direct ways to monitor the sender's true data, a planner can nevertheless design the underlying experiment to elicit more informative disclosures. For large-enough datasets, the quality of evidence an experiment provides to the receiver is contingent only on how well its most informative outcomes distinguish one state from another state. In many cases, this sharp fact of big data gives the designer a way to restrict to a lower-dimensional dataset, without loss of efficiency.

This work suggests several compelling directions for future research. One concerns voluntary disclosure with endogenous, costly data acquisition. For example, we would like to know how a sender might acquire data in order to persuade through voluntary disclosures. We conjecture that, if having more data benefits senders strategically regardless of its informational value, then data might be systematically over-collected precisely when it is cheap and plentiful. Another direction would consider the incentive effects of voluntary disclosure for agents: for example, some high-value ideas may see little investment because it is hard to distinguish success in realizing those ideas from success in other objectives based on evidence. Finally, there is a line of work that brings these questions closer to the ground by considering how they manifest in the structured data-generating processes commonly assumed in statistical or econometric models, for instance in logit and probit treatment effects models, or in linear models with Gaussian error. Doing so could help guide practitioners on what strategic omission looks like in those settings, and when it most undermines the value of research.

Chapter 2

Model (Non)-disclosure in Supervisory Stress Tests

2.1 Introduction

Every year, the Federal Reserve Bank performs a stress test on major banks and financial institutions in the US to assess the financial stability of the system under the Dodd-Frank Act. These stress tests are loss projections computed by the Fed given theoretical scenarios and the banks' balance sheet. The Federal Reserve Bank reveals parts, but not all, of the mapping ('model') from scenarios and balance sheets to loss projections, and this creates a tension in *model disclosure*, which this paper seeks to investigate.

Risk managers claim that not knowing the model parameters the Fed is using adds further risk to their decisions. On top of the underlying risk any financial action has, they must project how this will affect the stress test results. Under that view, banks do not know the regulatory cost of risk, and this regulatory uncertainty induces banks to make inefficient decisions, such as excessively reducing lending (Gissler et al. (2016)).

On the other hand, the regulators are concerned of potential dangers of revealing the stress test models. According to the Federal Reserve's memo,[1] there are three reasons to not keep the test model parameters and assumptions transparent:

1. *Gaming the system*: "Firms could [..] make modifications to their businesses that change the results of the stress test without changing the risks they face."
2. *Correlation*: "[Full disclosure] could increase correlation in asset holdings [..] making the financial system more vulnerable."

[1] https://www.govinfo.gov/content/pkg/FR-2017-12-15/pdf/2017-26856.pdf. Retrieved May 13, 2020.

3. *Model Monoculture*: "Full disclosure could incentivize banks to simply use models similar to the Federal Reserve's, rather than build their own capacity to identify, measure, and manage risk."

The Federal Reserve's concern on *gaming the system* and *model monoculture* motivate our stylized model. In our framework, both the Fed and banks have imperfect information ('models') about the underlying state regarding a risky asset, and banks take actions ('investments') which affect the Fed's payoffs. The Fed cares more about systemic risk or the possibility of a financial crisis than individual banks; we highlight this by assuming that the bank would always like to invest in a risky asset, while the regulator's utility ('social welfare') is such that the regulator wants the bank to invest only in good states. The Fed's private information corresponds to the Fed's model and information about the economy, which the Fed uses to 'test' the bank's investments under stress, and debates whether to release more or less information about.

In this setup, the trade-off is as follows. The Federal Reserve can give banks more information about the models they use for the test, which would help the banks make more efficient decisions. At the same time, giving more information about these models could incentivize banks to simply use those models to assess their actions, instead of using their own models. Under this trade-off, what is the optimal disclosure policy for the Fed?

In our stylized model, there is a regulator and a bank, each with private, imperfect information about the riskiness of an asset. The regulator can 'punish' banks for making investments in 'bad' assets determined by its own model, but can choose to disclose additional information about the punishment to affect the beliefs and the action of the bank. In a theoretical sense, our setup is a mixture of both optimal mechanism design and information design (a la Bergemann and Morris (2016a)): the principal can design punishments, as well as disclose information.

The model we propose concisely captures the trade-off in disclosing more information. Disclosing more information reduces the variance in the bank's decision problem, allowing a more informed decision; at the same time, disclosing more information allows the bank to optimize their action with respect to the regulator's punishment without using their own information. If the regulator fully discloses its model, banks already know everything about the punishment, and can expose themselves to (socially) excessive risk as long as they can pass the regulator's stress test. Facing such a trade-off, what is the optimal disclosure policy for the regulator?

Our answer is as follows. We first show that the regulator fully disclosing everything is not first-best, because the bank will *game the test*, investing in the risky asset if they can pass the test, even if their private information indicate significant risk. Then we show that even

when the regulator is restricted to a binary choice between full disclosure to no disclosure, not disclosing any information can outperform fully disclosing if the bank has more precise information than the regulator (Proposition 7). Next, if the regulator can choose arbitrary partial disclosure policies, the regulator can fine-tune a disclosure policy to approximate the first-best - make the bank choose the socially optimal action, incorporating both the regulator and the bank's private information, which is impossible in both full and no disclosure settings (Proposition 9). We discuss the implications of this proposition.

2.1.1 Relevant Literature

This paper contributes to a strand of literature in stress test design, especially on the question of *disclosure* in stress tests. Recent developments in the question of stress test disclosure include Goldstein and Leitner (2018) which study the disclosure problem of stress tests to agents with heterogeneous beliefs, Parlatore and Philippon (2018) which study the design of stress scenarios, Inostroza and Pavan (2020) which cast the coordination problem of receivers into a global game with private information, and Parlasca (2019) studying the time inconsistency problem in stress testing. The Handbook of Financial Stress Testing by Farmer et al. (2022) include several chapters dedicated to stress test design and disclosure, most notably Goldstein and Leitner (2022) which overview the literature of stress test disclosure.

Most of the papers in the literature of stress test disclosure concern the disclosure of stress test *results to investors* and its implications, after the stress test takes place. On the other hand, the disclosure problem and the trade-offs we study are regarding the disclosure of stress test *model to banks*. The paper that explicitly studies the question of *model disclosure* is Leitner and Williams (2022), which is the first paper to provide a unified framework to study disclosure of regulators' stress test models, which we build on. The main departure we have from their framework is that their paper implicitly assumes that the banks have an asymmetric informational advantage about the state: everything that the regulator knows is an obfuscation, or a *garbling* of what the banks know.

Instead, our framework considers the possibility that the banks may not be fully informed, or that the regulator may have some informational advantage over banks. We believe this is a more reasonable assumption that leads to different conclusions. Indeed, the Federal reserve does seem to have some informational advantages over individual banks: (1) macro-variables that the Fed would pay attention to, more than individual banks[2] (2) systemic risk that aggregates information reported from all banks; each bank knows its own position, but it's only the Fed that knows and takes into account each bank's position in the market. This is

[2]Former Boston Fed President Eric Rosengren: "We asked BofA to tell us their exposure to subprime mortgages.. they had no idea"

relevant to supervisory stress tests since any such informational advantage the Fed has will be used in conducting the stress tests.

If the banks are not fully informed, by releasing information about the underlying state, the regulator can provide *guidance* to help agents make a more informed decision, which can be beneficial for everyone. By excluding this effect, Leitner and Williams (2022) conclude the regulator should never fully disclose their information. We clarify in Proposition 7 that the regulator does not want to disclose if and only if it believes that the banks are more informed. In both papers, the regulator can do better by partially disclosing its information. However, we expand the disclosure policies in Section 2.4, show how the regulator can achieve the first best outcome with sufficient flexibility on the information structure: notably, we show that approximating the first-best outcome is possible by a disclosure structure that almost fully discloses the regulator's information.

From a theoretical standpoint, our modeling framework relates to information disclosure in mechanism design. Our setup can be considered as a partially informed principal having the ability to design incentives and choose the amount of private information it wants to disclose, to a partially informed agent. Supervisory stress tests have this characteristic and we use it as an application to contribute to this literature at large.

The question of whether a principal with private information should fully disclose any information she has is of considerable interest, yet the literature gives mixed answers. In auction environments, Milgrom and Weber (1982) and Ottaviani and Prat (2001) discover the linkage principle and argue that the seller has incentive to fully reveal any information related with the buyer's valuation. On the other hand, it is well known in the persuasion literature (Kamenica and Gentzkow (2011a), Bergemann and Morris (2016a), Bergemann and Morris (2016b)) that an information designer can benefit from not fully disclosing, and instead randomizing over information structures.

In standard Bayesian persuasion literature, the payoff structures are exogenously given. On the other hand, standard literature on principal-agent problems feature a principal given a fixed information structure and choosing the optimal contract satisfying incentive compatibility constraints. The question of our paper involves endogenizing both the payoffs and the information by allowing the principal to design mechanisms as well as a disclosure policy to induce a certain action. Thus this paper most closely relates to the literature on information disclosure in mechanism design where the principal controls the private information agents learn beyond their initial private information. Developments in this literature include Eso and Szentes (2007), Bergemann and Pesendorfer (2007), Li and Shi (2017), Yamashita (2018) and Krähmer (2020).

There are two main differences between our paper and the aforementioned literature on disclosure and mechanism design. First, the previous papers analyze the case where either the agent or the principal with private information, while our paper concerns the case where both the principal and the agent have private, imperfect information. This matters for two reasons: first, in our setup, the principal discloses her private, imperfect information about the underlying state, whereas previous work assumes that the principal discloses information about the underlying state itself: thus the meaning of full disclosure differs in their setting and ours. This distinction is critical, since our setup has a trivial solution once we assume the principal has perfect information. The second difference is that previous papers concern mechanism design environments with transfers from the agent to the principal, while we consider principal-agent problems where the principal can punish the agent, but the disutility to the agent from punishment does not translate into gains for the principal. We believe our assumptions are closer to reality in studying financial stress tests, where a regulator has imperfect information and wants to influence an agent's action using punishments based on this imperfect information.

Our paper highlights the principal's gains from garbling information by keeping the agent guessing using his own private information, which is correlated with the underlying state and the principal's information. This intuition is closely related to the literature on surplus extraction through correlation, building on the seminal result of Crémer and McLean (1988) which shows that for any (quasi-linear) utility function and a finite type space, if the agents' signals are correlated, there exists a mechanism that extracts all the surplus. McAfee and Reny (1992) prove this result in a continuum; Rahman (2012) generalizes this result to arbitrary type spaces. The main argument in Cremer-McLean is to use other bidders' bids that are correlated with each bidder's valuation to construct payoffs, to make each bidder effectively report their own value. In a similar vein, the main strategy of a principal with the power to arbitrarily disclose her information and design arbitrary punishments is to use her own information, which is correlated with the agent's information: by making the agent guess the principal's information, the principal can effectively elicit the agent's private information.

A contribution we make to the surplus extraction literature is providing a real-life application to the Cremer-McLean surplus extraction result. While a celebrated result in theory, applying Cremer-McLean's main insight of "taking advantage of correlations to extract information rent" in practice is difficult: Milgrom (2004) has described the full surplus extraction result as "nothing that is found in practice and reminds us of how important it is to check the practical reasonableness of solutions suggested by a model before implementing any practical policy based on the model." Our setup and argument gives a practical application of a principal taking advantage of correlation structures in information, by "keeping the agent guessing" about the principal's private information. Proposition 9 extends to suggest the optimality of partial disclosure and the advantages in using the correlation in information to the principal's benefit.

Our result is also relevant to the distinction between "omniscient persuasion" and "public persuasion" (Bergemann and Morris (2016b)). We show that with carefully designed information structures and incentives, public persuasion can allow the sender to achieve utility arbitrarily close to that from omniscient persuasion. In public persuasion setups, incentive compatibility constraints restrict the set of attainable payoffs for the principal. Our paper analyzes the case where the principal cannot elicit the agent's private information through standard contracts because the principal is tied to only making binary 'punishment' decisions. We show that in such circumstances, randomizing on the principal's own information allows her to elicit the agent's private information.

2.2 Model Setup

There is a bank and a regulator. The bank can choose between two actions: invest in a risky asset or a safe asset. While both the bank and the regulator know the payoff of the safe asset, both the bank and the regulator only have private, imperfect information about the payoff of the risky asset. The regulator can choose two objects of interest: a *stress test*, which is an assessment of the expected performance of the risky asset in a given scenario, conducted using what the regulator knows; and a *disclosure policy* about what the regulator knows. The formal model specifications are laid out below.

Agents and Fundamentals. The economy consists of a bank (B) and a regulator (Fed, F)[3]. The bank can choose between two actions: invest in a risky asset or a safe asset. The payoff from investing in the safe asset is known; we normalize it to zero for both the bank and the regulator. The payoff of the risky asset depends on the realization of a random variable $\omega \in \Omega \subseteq \mathbb{R}$. We call ω the **underlying state** of the risky asset. The bank's payoff from the risky asset is $u^B(\omega)$, which are the private gains of investing in the risky asset (relative to the safe asset). Assume $u^B(\omega)$ is bounded over Ω, so that utility obtainable in the best possible state is $\bar{u}^B < \infty$. The regulator's payoff is $u^F(\omega)$, which can be considered as the social welfare associated with the risky asset, which the regulator, as a constrained social planner, seeks to maximize.

There is a conflict of interest: the bank always prefers the risky asset to the safe asset, while the regulator regulator prefers the risky asset only if the underlying state is good (safe). Formally,

Assumption 2. *u^F is an increasing function of ω, and there is a ω_0 such that $u^F(\omega) \geq 0$*

[3]The model naturally extends to multiple banks; we keep it to one bank for clarity. We discuss this extension in Section 2.5.

iff $\omega \geq \omega_0$.

Assumption 3. *u^B is an increasing function of ω and $u^B(\omega) \geq 0$ for all ω.*

Assumptions 2 and 3 highlight the main conflict of interest: the bank wants to invest even in risky states, whereas the Fed wants the bank to invest only in good states. This is the only assumption we make – we don't make any structural or parametric assumptions about the payoff function.[4] The following example previews the types of situation we are interested in:

Example. Assume there is a financial crisis with probability p. Let R_g be the (known) gross return of a risky asset when there is no crisis, and ω be the bank's return of the asset when the financial crisis occurs. Let $u^B(\omega) = (1-p)R_g + p\omega - 1$, $u^F(\omega) = (1-p)R_g + p\omega - pL - 1$. The assumption $u^B(\omega) \geq 0$ highlights the fact that the banks want to invest even when the return in the bad state is low, indicating a higher willingness to take risk and/or moral hazard associated with the Fed put; whereas the regulator is more concerned about the macroeconomic costs, including having to bail out such banks, and externalities from financial crisis (Caballero Simsek 2013).

This example generalizes to any adequate interpretation to an uncertainty regarding a risky asset: ω could denote Value-at-Risk, beta, or quality outstanding bonds. Depending on the level of risk aversion, we can construct an appropriate utility function for both the regulator and the bank.

Information (Model). Both the regulator and the bank have their own private 'models' of ω. Formally, we assume that both the regulator and the bank have a common prior $\omega \sim f$ on the underlying state, and their private model of ω is an imperfect signal on ω.[5] The regulator's model of ω is represented by a signal $s^F \in \mathcal{S}^F \subset \mathbb{R}$ drawn from a distribution with CDF $F_{s^F}(\cdot|\omega)$ and density $f_{s^F}(\cdot|\omega)$, and the bank's model of ω is represented by a signal $s^B \in \mathcal{S}^B \subset \mathbb{R}$, drawn from a distribution with CDF $F_{s^B}(\cdot|\omega)$ and density $f_{s^B}(\cdot|\omega)$. A natural interpretation of s^F, s^B is model-implied values of ω.

Here the prior f captures any common knowledge between the regulator and the bank, while s^F, s^B capture what the regulator and the bank respectively know about the underlying state, represented in their models. We assume that the bank observes s^B and the regulator observes s^F before the bank makes its investment decision (hence a "model" of the risky asset

[4]The assumption that $u^B(\omega) \geq 0$ for all ω highlight the bank's stronger desire to invest, or some form of moral hazard in the banks' belief that they will be rescued by a bailout when the underlying state turns out to be bad (such as a financial crisis). It also helps simplify our analysis.

[5]The common prior assumption is not essential; the regulator and bank can agree to disagree on a different prior. As long as the heterogeneous prior is common knowledge, our analysis follows through.

before the returns are realized). We call s^F (resp. s^B) the regulator (resp. bank)'s information or signal. In practice, the FED model to asses losses from bank's positions is comprised of several regression equations per loan proftolio with unknown coefficients, estimated with previous data.

We do not impose parametric forms on the distributions, but we assume that both the regulator and bank are more likely to observe higher signals when the state ω is higher:

Assumption 4. *$s^F|\omega$ and $s^B|\omega$ satisfy monotone likelihood ratio property (MLRP): for any $\omega_1 > \omega_2$, both $f_{s^F}(s|\omega_1)/f_{s^F}(s|\omega_2)$ and $f_{s^B}(s|\omega_1)/f_{s^B}(s|\omega_2)$ are strictly increasing in s.*

This ensures that both the regulator and the bank is more likely to observe higher signals when the underlying state is larger; moreover, if the bank sees a higher signal, the regulator's signal is likely to be higher too (and vice versa).

As we pointed out in the introduction, this formulation implicitly assumes that the Fed's model of ω contains some information *of intrinsic value* that the bank does not know (captured by s^F). This departs from previous work such as Leitner and Williams (2020) where the bank's private information is clearly superior to the regulator's[6]. We believe our assumption that the regulator and banks independently have some nonoverlapping private information is realistic, and delivers crisp implications based on the relative 'quality' of information and models that the regulator and banks may have.

Stress test. After the bank makes its investment decision, the regulator conducts a supervisory stress test to determine whether to pass or fail the bank. Investments in safe assets always pass the test. If the bank invests in the risky asset, the regulator uses its private information to 'test' the investment. Specifically, the bank's investment passes the regulator's stress test if and only if the regulator's private information s^F is above some threshold s^*. In practice, this threshold could represent model-implied threshold Value-at-Risk in adverse scenarios, or minimum capital requirements. The regulator chooses and announces this threshold s^* *before* the bank and Fed observe s^F, s^B.

Assigning a threshold rule on s^F for supervisory stress tests highlight the idea that supervisory stress tests are the regulator testing whether the risky investment is safe enough under an adverse scenario, using its own models and parameters, as opposed to using the bank's own assessments of their investments. Indeed, s^F encapsulates all of the regulator's private information that it uses to estimate or evaluate ω. Thus it is natural to interpret s^F as an output of the regulator's model it uses to test the bank's risky investment, or a key coefficient/parameter in the regulator's model.[7]

[6] In Leitner and Williams (2020), the regulator's information is a strict garbling of the bank's information.

[7] The debate around disclosure of the Federal Reserve's Dodd-Frank Act Stress Tests is centered on

We can interpret s^* as the **strictness** of the test – with a higher s^*, the regulator is less likely to give a pass to a risky investment, so the test is stricter.

Failure. If the bank invests in the risky asset and fails the test, a private cost c is incurred to the bank. We call (s^*, c) the *stress test structure*.

Multiple interpretations for this cost exist in this context – one is a cost to the bank from "failure" being publicly announced to the market; this would lead to a decline in the market's confidence to the bank, incurring costs such as stock price declines (Flannery, Hirtle, Kovner). Another is direct transaction costs associated with having to liquidate assets. Yet another can be a direct "punishment" imposed by the Fed, such as restrictions on share repurchases or dividends, cease-and-desist orders to correct practices in risk management, or even penalty fees. All of these punishments have been used by the Federal Reserve in conducting the Dodd-Frank Act Stress Test (DFAST).

Note that the aforementioned interpretations are ambiguous as to whether the regulator may have control of c. It seems that under some interpretations (direct punishments/regulations or intensity of signalling to the market), the Fed may have some control over the magnitude of the cost, whereas it seems too strong to assume that the Fed can fine-tune c. In this paper, We explore the case when c is exogenous, and discuss the case where the Fed can set c optimally.

We can interpret c as the **harshness** of the test: a higher c implies that the costs of failing the test are harsher on banks.

Final payoffs. After the stress test and punishments, the payoffs are as follows. If the bank invests in the safe asset, the regulator and the bank's payoffs are 0. If the bank invests in the risky asset and passes the stress test, it receives payoff $u^B(\omega)$ and the regulator receives payoff $u^F(\omega)$. If the bank invests in the risky asset and fails the stress test, it receives payoff $u^B(\omega) - c$ and the regulator receives payoff $u^F(\omega)$.

Disclosure. The main question we seek to answer is whether the regulator should reveal its "model" s^F to the bank. We analyze this in a Bayesian persuasion framework as in Kamenica and Gentzkow (2011): the regulator can *partially disclose* information about s^F by committing to an information structure that provides the bank with signals about s^F. In the supervisory stress test framework, this corresponds to giving some, but not all, of the

revealing the coefficients of regressions that the Fed uses to estimate losses.

parameters and model specifications that the Fed uses in estimating ω.

Formally, before observing s^F, the regulator discloses information about s^F by committing to a *disclosure policy* (M, π) that consists of a set M of messages and conditional distributions $\pi : \mathcal{S}^F \to \Delta(M)$, where $\pi_{s^F} \in \Delta(M)$ denotes the distribution of messages conditional on s^F. An equivalent way to state the definition of π is as follows: after observing s^F, the regulator generates a *message* m about s^F according to $m|s^F \sim \pi_{s^F}$ and reveal m to the bank. This message m captures (partial) information about the regulator's model that it discloses to the banks.

Two benchmark information structures are *full disclosure* corresponding to $\pi_{s^F}^{full} = \delta_{s^F}$, the point mass on s^F, and *no disclosure* corresponding to $\pi_{s^F}^{no} = \delta_m$, the point mass on some m independent of s^F: the regulator sends the same message m for any s^F, disclosing no information about s^F.

Timing. The timing of the model is as follows:

1. (Ex-ante stage) The regulator chooses a disclosure policy π about s^F and a punishment threshold s^*, and publicly announces it.
2. (Interim stage) The regulator observes s^F, and the bank observes s^B.
3. (Disclosure stage) The regulator discloses $m \sim \pi_{s^F}$ to the bank, and the bank updates its posterior on ω and s^F.
4. Knowing s^B and m, the bank decides whether or not to invest in the risky asset.
5. If the bank decides to invest and $s^F < s^*$, the bank fails the test.

Problem. The regulator and the bank's problem is given as follows.

- **Bank's problem.** The bank decides whether or not to invest in the risky asset, knowing the bank's own signal s^B and the regulator's message m. Given s^B, m, the bank believes investing in the risky asset will *pass* the stress test with probability $q = P[s^F > s^*|s^B, m]$. Then the bank invests if and only if

$$E[u^B(\omega)|s^B, m] - (1-q)c \geq 0 \tag{2.1}$$

  Note that q is increasing in s^B, and $u^B(\omega)$ is monotonic in ω, so the expectation is increasing in s^B. Thus given m, the bank invests if and only if $s^B \geq s^{B*}(m, c)$ for some s^{B*}.

- **Regulator's problem.** The regulator chooses s^* (stress test threshold), pi (disclosure policy) in the ex-ante stage (before observing s^F) to maximize

$$\mathbb{E}_{\omega, s^F, s^B, m}[u^F(\omega)] \tag{2.2}$$

  where the expectation is over all possible realizations of ω, s^F, s^B and m.

The actual realization of ω is irrelevant to the regulator or the bank's decision problem, as the decision happens before the realization. This is natural in our context: any decision related to investment in risky assets, including supervisory stress tests and punishments, cannot depend on the actual payoff of the risky asset. We define the equilibrium as:

Definition 2.2.1. *(Equilibrium) An equilibrium is a disclosure policy $\pi : \mathcal{S}^F \to \Delta(M)$, stress test threshold s^*, and action chosen by the bank (invest in safe or risky asset) such that the regulator and bank each solve their optimization problem.*

First-best action for the regulator. Given information μ about ω, the regulator wants the bank to invest in the risky asset if and only if $E[u^F(\omega)|\mu] \geq 0$. There are two sets of information available in the market: s^F known by the regulator and s^B known by the bank. Thus the first-best action is to invest iff

$$E[u^F(\omega)|s^F, s^B] \geq 0.$$

This takes into account both the information that the regulator and bank knows. Note that the expectation is increasing in both s^F and s^B, by Assumption 4.

Call the set of signals $S^{FB} = \{(s^F, s^B)|E[u^F(\omega)|s^F, s^B] \geq 0\}$; this is the set of *realizations of information* for which the regulator wants the bank to invest. Also denote by $U^F(s^F, s^B) = E[u^F(R)|s^F, s^B]$ the expected utility of the regulator when the bank invests in the risky asset, when the Fed knows s^F and the bank knows s^B. We seek to answer the following questions:
Research Question. Should the regulator reveal s^F to the bank prior to the investment? Can hiding, or garbling and partially revealing the regulator's model/information be welfare enhancing than fully disclosing it? How strict should the stress test be, and how does this depend on the fundamentals?

2.3 Full vs. No Disclosure when banks want to pass

In this section, we compare between two disclosure policies: full disclosure where the regulator reveals s^F fully to the bank, and no disclosure where the regulator reveals no information about s^F to the bank. Restricting the regulator to this binary choice allows us to concisely deliver the intuition that disclosure may hurt the regulator because the banks would rely less on their own private information in investment decisions and just follow the supervisory stress tests. Moreover, this is not a gross simplification – if the regulator cannot credibly commit to more complicated disclosure policies, the only choices would be to reveal or not reveal.

For simplicity, we assume that banks always want to pass the stress test:

Assumption 5. *There is some $c_0 > \bar{u}_B$ that lower bounds the minimum possible punishment that the Fed can carry out conditional on failure.*

Since punishment outweighs any possible benefit to the bank from investing, banks will never invest in an asset if they are certain it will result in stress test failure.

2.3.1 Full Disclosure

When the Fed fully discloses s^F, the bank knows exactly when it will pass or fail the test. Thus the expected utility from investing in the risky asset is

$$U^{B,FD}(s^B, s^F) = \begin{cases} E[u^B(\omega)|s^B, s^F] > 0 & \text{if } s^F \geq s^* \\ E[u^B(\omega)|s^B, s^F] - c < 0 & \text{if } s^F < s^* \end{cases}$$

As mentioned before, the cost of failing the stress test is assumed to be high, i.e. $c \geq E[u^B(\omega)|s^B]$ for any s^B. The bank will invest if and only if $s^F \geq s^*$, i.e. if it knows the Fed's signal is high enough that it passes the test. The regulator's ex-ante utility from full disclosure with threshold s^* is

$$U^{F,FD}(s^*) = E[U^F(s^F, s^B)|s^F > s^*]Pr(s^F > s^*) = \int_{s^F \geq s^*} U^F(s^F, s^B)dF(s^F)$$

The optimal threshold s^{*FD} is the unique solution to $E[U^F(s^F, s^B)|s^F = s^{*FD}] = 0$. The reason why full disclosure cannot achieve the first-best for the regulator is clear; the bank would always invest if they knew they would pass the test, even if their information s^B suggests that the social welfare resulting from the investment is low. This is to say, the bank "games the system" when the Fed's signal is high enough to allow them to comfortably pass.

2.3.2 No disclosure

On the other hand, assume that the regulator discloses nothing about s^F. The bank's utility from investing in the risky asset is

$$U^{B,ND}(s^B) = E[u^B(\omega)|s^B] - (1-q)c$$

where $q = P[s^F \geq s^*|s^B]$ is now the bank's estimate of the probability of passing the stress test: the bank needs to "guess" the test results using its information s^B. Since q and $E[u^B(\omega)|s^B]$ are both increasing in s^B, $U^{B,ND}(s^B)$ is increasing in s^B, there is $s^{B*}(s^*, c)$ such that bank will invest iff $s^B \geq s^{B*}(s^*, c)$.

Thus the set of signals where the bank invests in the risky asset and passes the test is given by

$$S^{ND}(s^*) = \{(s^F, s^B)|s^B \geq s^{B*}(s^*, c)\}.$$

The mapping $s^* \to s^{B*}(s^*, c)$ is increasing in s^* and c; if the test is stricter or harsher, the bank would need a higher signal to invest. Thus $s^{B*}(s^*, c)$ is invertible with respect to s^*, and for a fixed c, we can define $s^*(s^{B*})$ be the value of s^* that implements s^{B*}

We let

$$U^{F,ND}(s^*) = E[U^F(s^F, s^B)|s^B \geq s^{B*}(s^*, c)]$$

be the regulator's payoff when it chooses the bank's threshold s^{B*}. Let $U^{ND*} = \max_{s^{B*}} U^{F,ND}(s^{B*})$ be the highest payoff to the regulator. The optimal policy satisfies $E[U^F(s^F, s^B)|s^B = s^{B*}(s^*, c)] = 0$.

The reason why no disclosure cannot achieve the first-best for the regulator is different now; as the bank does not know s^F, the bank is not fully informed. This clarifies the trade-off with disclosure: full disclosure makes the banks more informed, but at the same time, opens the door for the banks to *game the system*, investing even when the bank's assessment is bad (s^F high, s^B low).

Comparative statics

The payoff to the Fed under the optimal full-disclosure threshold, $U^{*FD} = U^{F,FD}(s^{*FD})$, is unaffected by the distribution of the bank's signal, and increases with the informativeness of the Fed's signal. To see the former, observe that the bank's decision is independent of its own signal. Note, for the latter, that if a signal $s^{F'}$ is Blackwell more informative than s^F, then $s^F = s^{F'} + \epsilon$ where ϵ is a noise term distributed $F_\epsilon(\epsilon|s^{F'})$.[8] If the optimal threshold under signal structure s^F is s^*, then under $s^{F'}$ the regulator could mimic expected outcomes under s^F by randomizing whether to allow banks to invest, and using an interior probability

$$p^F(s^{F'}) = 1 - F_\epsilon(s^* - s^{F'}|s^{F'})$$

of announcing to banks that they will be allowed to pass the stress test even if they invest in the risky asset. However, because there is a uniquely optimal cutoff under $s^{F'}$, and elsewhere it is either strictly optimal to invest or strictly optimal not to invest given s^F, randomization is suboptimal. The Fed therefore does strictly better by passing banks above the cutoff and failing them below, which it can enact when playing optimally with the more-informative signal $s^{F'}$, but not under the less-informative s^F.

On the other hand, the Fed's payoff under the no-disclosure threshold depends only on the informativeness of the bank's signal. Here, a more informative signal for the bank is better. The argument is similar to the previous case: since s^* determines the outcome through s^{B*}, the implied cutoff in the bank's private signal that determines whether they invest or not,

[8] We assume that the garbling preserves monotonicity of returns to investing in s^F; if not, then a cutoff rule may no longer be optimal.

and $U^{F,ND}$ depends only on s^{B*} and the joint distribution of s^B and ω. If $s^B = s^{B'} + \epsilon$ with $\epsilon \sim F(\epsilon|s^{B'})$, then the outcome of the optimal policy s^{B*} under s^B is equivalent to having the bank that observes $s^{B'}$ randomize, and invest with probability

$$p^B(s^{B'}) = 1 - F_\epsilon(s^{B*} - s^{B'}|s^{B'}).$$

Again, because the Fed's expected payoff from investment is also strictly increasing in the bank's signal, randomization is suboptimal everywhere except the optimal cutoff, and it is better for the Fed if the bank plays their best response to $s^{*\prime}$ using their more informative signal $s^{B'}$.

2.3.3 Full disclosure vs no disclosure

Which of the two regimes is preferred? The benefit of full disclosure is that the bank is more informed in its decision about ω, but it comes at a cost – namely, the bank *overinvests* in states where it knows it'll pass the test, and may *underinvest* when it knows it'll fail the test. On the other hand, without disclosure, the bank's investment decision is made only using its information, so some information is lost.

The comparative statics in the previous section show that full disclosure improves when the Fed has a precise signal, while no disclosure improves when the bank's signal is more informative. It follows that which regime is better depends on "whose model is better", i.e., on the relative precision of s^F and s^B.

Proposition 7. *Suppose the regulator chooses between full disclosure (FD) and no disclosure (ND), and c is exogenously given and fixed.*

1. *Fix s_F. If $U^{FD*} > U^{ND*}$ under s^B, then the same is true for all $s^{B'}$ that are garblings of s^B. In other words, full disclosure becomes more attractive to the Fed as the bank's model worsens.*
2. *Fix s_B. If $U^{FD*} > U^{ND*}$ under s^F, then the same is true for all $s^{F'}$ that are Blackwell more informative than s^B – when the Fed's own model improves, full disclosure becomes more attractive.*
3. *If s_F is a garbling of s_B, then no disclosure is better than full disclosure. If s_B is a garbling of s^F, then full disclosure outperforms no disclosure.*

Essentially, the Fed would like to disclose its model ONLY if its information is precise enough so that the Fed benefits from dictating the investment. Otherwise, it should hide its model, in order to prevent banks, which have more valuable information about tail-end risks, from discarding their own predictions, and to make banks leverage their own information in the Fed's favor by guessing the Fed's model.[9]

[9]The intuition resembles that of Cremer and McLean (1985): as long as the Fed's signal is not fully known

Proposition 7 clarifies to us that if there's any motivation for the regulator to disclose information, it is not because the regulator believes the banks are better equipped to make investment decisions and wants to delegate the decision to the bank. Rather, the regulator would like to fully disclose only if it believes that its own models are more accurate than the banks, to *provide guidance*, or make binding recommendations on investment decisions based on the regulator's model. This may have been true in the first few years following the Great Recession, when there was good reason to believe that banks themselves did not have sufficient capacity to identify systemic risk.

On the other hand, in more recent times, where it is generally believed that banks employ more sophisticated models and may have some proprietary information that allows them to have a more precise estimate of risk, the fact that banks have better models may actually incentivize the regulator to *hide* information, precisely because the regulator does not want banks to overinvest in risky assets that pass the regulator's test. In our framework, it is precisely because the regulator believes banks may have better private information that they hide the regulator's information, as the regulator wants to align the bank's interests with social welfare, while ensuring that the bank uses its private information.

2.3.4 Comparative statics on strictness and harshness

We investigate some comparative statics on fundamentals of interest – such as the passing threshold s^* ('strictness'), cost of failure c ('harshness'), and how they relate to the disclosure policy. We test predictions such as the following:

1. If the regulator discloses information, the regulator would have to make the tests stricter ("minimum required capital levels would need to be materially increased") to counteract gaming. This was suggested by former Fed governor Tarullo (2017).[10]
2. If the cost of punishment is higher to the banks, the test should be less strict.

We first show that the first prediction need not necessarily be true. Formally, if s^{*FD} is the regulator's optimal threshold under full disclosure, and $s^{*ND}(c)$ is the regulator's optimal threshold under no disclosure, the statement is equivalent to $s^{FD*} > s^{*ND}(c)$. However, as we show in the appendix, this is not necessarily true: it depends on the specific information structure, and what investment threshold the regulator wants to induce the bank to use in the case of no disclosure.

On the other hand, we have seen from the analysis of the full and no disclosure case that the second proposition is true for no disclosure, but not for full disclosure. Indeed, under no disclosure, a higher c is accompanied by a lower s^*. That is, if the cost of failure is

to the bank, it is correlated with the bank's own model, and this correlation can be exploited to make the bank use its private information in a way that best benefits the Fed.

[10] See: https://www.federalreserve.gov/newsevents/speech/tarullo20170404a.htm

harsher, the regulator responds by making the test *less strict*; if the cost of failure is lower, the regulator makes the test harder to pass. On the other hand, under full disclosure, as long as the cost of failing the stress test is high enough so that no bank would voluntarily fail the test with probability 1 (Assumption 5), the optimal threshold s^{*FD} is simply given by $E[U^F(s^F, s^B)|s^F = s^{*FD}] = 0$. As such, the trade-off between strictness and harshness exists only when the regulator hides information about the test.

2.3.5 Punishment flexibility

In Section 2.2, we discussed the possibility that the regulator may be able to *choose* the harshness of punishment c, through either sending stronger messages about the investment's riskiness, or directly choosing which punishment to levy (freezing / coercing to sell assets, fining for failure etc.) What happens if the regulator chooses c optimally, in either full disclosure or no disclosure? We briefly discuss the intuition here.

Under full disclosure, as we have seen in the above section, as long as the punishment c is large enough, the bank's investment decision solely depends on s^*, the threshold set by the regulator; banks will invest if and only if they know they will pass the test, as no bank will *intentionally* fail the test. Thus under Assumption 5, there is no gains from punishment flexibility. Instead, if we allow the regulator to set punishment c low so that *some banks may intentionally fail the test and invest in risky assets*, this may strictly benefit the regulator, because the banks who would be willing to take the risk are selected to be the ones whose models implied a *better state*; so the regulator's utility can improve if the regulator allows some banks to fail. However, allowing some bavnks to intentionally fail is unrealistic – most notably, market participants learning failure is a strong enough disincentive sets a lower bouind for c, so it is likely that many banks to not want to intentionally fail the test.

Under no disclosure, choosing c and choosing s^* are *dual* in the sense that both are tools to affect the bank's investment threshold $s^{B*}(c, s^*)$; we have seen above that a higher c (harsher) is associated with a lower s^* (less strict). As such, the ability to choose c does not affect the optimal policy. Thus, under both full disclosure and no disclosure, the regulator's capacity to choose c does not significantly affect our analysis.

2.4 Partial Disclosure

We now assess the optimality of *partial disclosure* policies, where the regulator can commit to providing some information about its private information about the state of the world. In this setup, the regulator, upon observing its private signal s^F, can choose to disclose partial information about s^F, without fully disclosing it. As mentioned in Section 2.2, we model this as the regulator choosing a disclosure policy (M, π) *before observing* s^F such that, upon observing s^F, the bank discloses $m \sim \pi_{s^F}$. The bank updates its posterior of the state ω and

the regulator's private signal s^F using m and makes its investment decision.

Under this 'middle ground' between full disclosure and no disclosure, how much better can the regulator perform? We argue here that fine-tuning the disclosure policy is sufficient to be arbitrarily close to the first-best, as long as the harshness c is sufficiently large.

2.4.1 The bank and the regulator's problem

Suppose regulator chooses the strictness s^* and the disclosure policy (M, π). Upon seeing s^F, the Fed discloses $m \sim \pi_{s^F}$. The bank, upon observing m, will form posteriors on ω and s^F given s^B and m: $\hat{\omega} \sim (\omega|s^B, m)$ and $\hat{s}^F \sim (s^F|s^B, m)$. Given this information, the bank will invest if and only if

$$U^B(s^B, m) = E[u^B(\omega)|s^B, m] - (1 - P[s^F > s^*|s^B, m])c \geq 0$$

This expected utility is increasing in s^B. As such, the bank invests if and only if $s^B \geq s^{B*}(m, c)$ for some mapping s^{B*}. Since the utility is decreasing in c, this mapping s^{B*} is clearly increasing in c; the banks need a higher signal to invest in a risky asset when the punishment is harsher. The dependence on m is trickier because we have no guarantee that m is monotonic - for example, the regulator may want to pool extremely good signals with extremely bad signals. But if we rule out such disclosure policies, we have a monotonicity result:

Lemma 8. *Suppose that the regulator's disclosure policy $m|s^F$ satisfies MLRP: for $s_1^F > s_2^F$, $f(m|s_1^F)/f(m|s_2^F)$ is increasing in m. Then the bank's expected utility $U^B(s^B, m)$ is increasing in m, so $s^{B*}(m, c)$ is decreasing in m.*

The regulator needs to construct a disclosure policy (M, π) and s^* to maximize ex-ante utility

$$\mathbb{E}_{\omega, s^F, s^B, m}[u^F(\omega)]$$

Recall that the first-best action that the regulator would like the bank to take is to invest in the risky asset iff

$$\mathbb{E}[u^F(\omega)|s^F, s^B] \geq 0$$

Since $s^F|\omega$ and $s^B|\omega$ both satisfy MLRP, the left-hand side is increasing in s^F, s^B. So there exists some decreasing $s^{B,FB}(s^F)$ such that the first-best action is for the bank to invest if and only if $s^B \geq s^{B,FB}(s^F)$.

From this formulation we clearly see the trade-offs of disclosure. Under full disclosure, the bank's decision rule $s^{B*} \geq s^{B*}(m, c)$ will depend exclusively on $m = s^F$, and not on s^B. With less disclosure, however, the bank's decision rule $s^{B*} \geq s^{B*}(m, c)$ will be noisy compared to the first-best $s^{B,FB}(s^F)$ since m is not s^F, and when the noise is sufficiently large, the banks will just rely on s^B to make the decision.

2.4.2 Approximating the first-best

The natural question is: can the regulator choose disclosure policy (M, π) and stress test s^* such that $s^B \geq s^{B,FB}(s^F)$ and $s^B \geq s^{B*}(m, c)$ are identical, or close enough? In this subsection, we show that if c is sufficiently large (or if the regulator can choose c), the regulator can design a stress test such that the regulator can induce the banks to approximate the first-best action that incorporates the signals of both the regulator and the bank. Thus this is clearly a strict improvement over full disclosure – when the bank can game the test – and over no disclosure – when the bank's action depends only on its signal.

Proposition 9. *Suppose the regulator can choose any information structure to partially disclose its signal. Then the regulator can design a disclosure policy with infinitesimal noise to approximate the "efficient investment" decision that incorporates both the information known by the regulator and the bank. Formally, for any $\epsilon > 0$, the Fed can choose the stress test (s^*, c) and a disclosure policy π such that the bank, after observing m will invest in the risky asset if and only if $s^B \geq s^{B*}(m, c)$, and this action satisfies*

$$E[u^F(\omega) 1_{s^B \geq s^{B*}}] > U^{FB} - \varepsilon$$

where U^{FB} is the first-best utility of the regulator, and the expectation on the left-hand side is taken over all signal and message realizations.

An intuition for the proof of the proposition is as follows. For each bank with a signal s^B, the regulator sets up a disclosure policy such that the regulator almost fully discloses their model s^F, and almost all of the banks pass the test, but there is always a small probability of failing the test, and the conditional probability of failure is larger when the bank sees a lower message ('higher probability that the Fed believes the investment is risky'). The regulator fine-tunes the disclosure policy (M, π) such that the bank's 'threshold signal' $s^{B*}(m)$ after seeing the regulator's message m closely approximates the welfare-maximizing investment threshold.

Specifically, the regulator chooses the disclosure policy (M, π), cost of failure c, and passing threshold s^* such that:

- $M = \mathcal{S}^F$: the set of messages is simply the set of private signals of the regulator.
- $\pi_{s^F} = (1 - \epsilon(s^F))\delta_{s^F} + \epsilon_{s^F} U(\mathcal{S}^F)$: the disclosure policy fully reveals s^F with probability $1 - \epsilon(s^F)$, but the message may be garbled with some noise with full support, so that the banks can never be 100% sure they will pass the test.
- Construct the 'noise probabilities' $\epsilon(s^F)$ to be small enough so that the disclosed message m is almost always equal to s^F, but upon seeing any message m, there is a probability $p(m, s^B)$ such that the bank fails the test.
- Set c high, and fine-tune the probabilities $p(m, s^B)$ (using $\epsilon(s^F)$) such that $s^{B*}(m, c)$ is sufficiently close to $s^{B,FB}(m)$, which is sufficiently close to $s^{B,FB}(s^F)$ as long as $\epsilon(s^F)$

is small enough.

A rigorous construction and proof of approximate optimality is given in the appendix.

This result highlights the main benefit of partially disclosing. The regulator wants the bank to 'guess' the regulator's signal s^F using s^B, because guessing elicits the bank's private information; however, hiding information is inherently costly as it creates additional noise, when the regulator's goal is to match the bank's actions with the information that both the regulator and banks know. But by disclosing almost everything but leaving a small noise, the regulator can keep the benefits of 'not disclosing' – eliciting the bank's private information and preventing gaming the system – while keeping almost all of the benefits from disclosing – allowing banks to make a more informed decision in investments.

This result too, is in line with the Federal Reserve's actual policy regarding disclosure in the Dodd-Frank Act Stress Test (DFAST). The Fed is increasingly disclosing more information because it "helps financial institutions [..] understand the capital implications of changes to their business activities, such as acquiring or selling a portfolio of assets," at the same time cautioning against full disclosure because "doing so could permit firms to reverse-engineer the stress test." Disclosing most, but not all, of the relevant models of the stress test allows the regulator to reap most of the benefits associated with disclosing, while keeping enough uncertainty to shut down the perils associated with disclosure.

In the limit when our disclosure policy approximates the first-best ($\varepsilon \to 0$), we have that: banks almost fully know the regulator's information ($P[m = s^F] \to 1$), the banks are punished harshly ($c \to \infty$), but the test is not strict ($P[s^F \leq s^*] \to 0$), and the bank's investment decision, conditional on the message, is the first-best investment decision. Our result is suggestive that the regulator disclosing more information is associated with a harsher test (Section 2.3.4), and if the regulator can choose between strictness and harshness (Section 2.3.5), it is better to make the test harsh (punish severely) but not strict (punish only a few banks); however, since our result shows one way to approximate first-best but doesn't show that it is the *only* way, other disclosure policies and punishments may be able to approximate the first-best as well.

However, there are two caveats in applying the insights from this proposition in reality:

1. The optimal disclosure policy may be difficult to implement in practice. To approximate the first-best, the regulator has to add an infinitesimally small probability of 'failure,' and fine-tune the probabilities so that the bank with signal s^B finds it exactly indifferent between investing and not investing at $s^B = s^{B*}(m, c)$. While the Federal Reserve definitely has some signaling capacity, it may be unrealistic to assume that the Federal Reserve can fine-tune the message to this extent.

2. The possibility of misspecification makes our result less robust. Our first-best approximation result assumes that the regulator is fully aware of the information acquisition structure of itself and the bank, and the banks fully know and believe the complicated disclosure policy and updates accurately in a Bayesian manner. If the regulator is misspecified in any of these steps, combining a small noise with large punishments could lead to very inefficient outcomes.

As such, Proposition 9 should not be taken literally, and should be considered as a benchmark that highlights the substance – hiding *some* information is always better than revealing all information, whereas hiding too much information is going to hurt social welfare. At the same time, the intuition the proposition suggests – as long as the regulator does not disclose its private information fully, more disclosure could be socially beneficial if it is accommodated by an appropriate punishment – is quite relevant in the stress test context.

2.5 Discussion and extensions

In this section, we discuss the implications of our main propositions and their intuition, and discuss possible extensions of our framework.

1. What if there is a social cost of banks failing the stress test that the regulator must account for? We assumed for simplicity that the regulator does not internalize the cost of punishment that is incurred to bank as a result of failing the stress test. If such social welfare costs of punishment exist, no disclosure will be relatively worse than full disclosure compared to our initial analysis, as under no disclosure some banks may still invest and get punished, while under full disclosure no bank will ever voluntarily invest knowing it will be punished. However, if the social cost is not as large as the private costs to banks (which include management changes, ban on additional transactions, etc.), it may still be the case that no disclosure outperforms full disclosure. Moreover, partial disclosure strictly dominates both full and no disclosure.

2. A natural extension would be to add multiple banks which make independent investment decisions. If the regulator has multiple banks, each with their own 'models' s_i^B, then the disclosure problem becomes whether or not to disclose a common s^F to multiple banks with their own s_i^B. In such a scenario, we immediately see that there is stronger incentive for the regulator to not fully disclose, as each bank using their own information would likely lead to a better outcome than every bank blindly investing the maximum according to the stress test rule. This can be formalized if the regulator has a specific aversion to banks' actions being correlated – either investment is multidimensional and there is some correlation disutility, or the regulator's social welfare function dislikes banks' investment thresholds being too close to one another. This highlights the concern expressed by the Federal Reserve against disclosing: "[Full disclosure] could increase correlation in asset holdings [..] making the financial system more vulnerable."

3. In a similar vein, to make our stylized model more realistic, we may treat investment and 'models' as multidimensional objects. This would clarify the 'information' in the disclosure problem as parameters in question as specific coefficients into the regulator's 'stress test' model, and make explicit some of the goals of the regulator (decrease the correlation across banks – low beta) versus the bank (increase correlation with market – high beta). In this case, there's an additional layer of optimal disclosure that can be discussed – the optimal number of dimensions that the regulator may disclose. This can be naturally interpreted as the number of coefficients to disclose in actual stress tests conducted by the Federal Reserve.

4. Endogenizing information acquisition by the banks. Another existing concern by the Federal Reserve argument against model disclosure is that "full disclosure could incentivize banks to simply use models similar to the Federal Reserve's, rather than build their own capacity to identify, measure, and manage risk." What if banks' draw s^B is not drawn from an exogenous distribution, but banks could invest in knowing a more precise s^B, through some information acquisition that is more costly in the precision of the draw? If the regulator fully discloses, then banks will never invest in knowing a better s^B. Hiding the regulator's model s^F would incentivize banks to "do their own research" to pass the stress test, hence investing in their own, independent models, and therefore also improves the outcomes of less relative to more disclosure. This allows room for investigating the optimal level of disclosure, if banks investing in their capacity comes with a social cost.

5. The paper focuses its analysis on supervisory stress tests in the financial sector, but the intuition and logic naturally extends to any environment where a regulator has imperfect information about an underlying state and wants to influence an agent's action by assessing and punishing agents' actions through her imperfect information. Applying our analysis to scenarios such as traffic cameras, law enforcement, or firm-employee relationships can explain why intentional obfuscations/garblings exist in many principal-agent relationships with private information.[11]

2.6 Conclusion

Should the Federal Reserve disclose its stress test models to banks? To shed light on this question, we propose a stylized model where a regulator and bank has private information about the state of a risky asset, and the regulator uses its own information to test the investment. Our framework incorporates the main trade-off of disclosure: disclosure allows banks to make a more informed decision, but at the same time, allows banks to *game the test*.

Our main contribution is highlighting the possibility of the regulator to elicit the banks'

[11] A previous version of this paper titled 'Optimal Disclosure in Principal-Agent Problems with Imperfect Information' discusses this in more detail.

private information by hiding the regulator's private information used for the stress tests, and keeping the banks guessing about this information. Our first result comparing full disclosure to no disclosure highlights this intuition: we show that no disclosure can be better than full disclosure if banks hold better information.

Our second result shows that if the regulator can commit to arbitrarily complex disclosure policies, the regulator can choose a disclosure policy that reveals almost everything, but leaves a small noise, and fine-tune this so that the bank's investment action is arbitrarily close to the socially optimal investment decision, and combining this with a sufficiently harsh punishment.

The second result may shed some light on the Federal Reserve's strategies in recent years, where they release more information on the crucial parameters and modeling structures, but keep some parts opaque, and punishment for failing the test is severe enough such that no major bank voluntarily fails the test. However, at the same time, the second result relies on fine-tuning information disclosure and committing to them; it should be thought of as a benchmark, and not too literally. When the regulator lacks commitment power, it may make more sense to think of the regulator's question as choosing between full disclosure and no disclosure, whence we are back to our initial framework in which the signals' relative informativeness matters.

While we lay out our intuition for the main trade-off of gaming the system, our model is stylized in nature, and cannot answer quantitative policy questions on stress test disclosure by itself. A natural next step will be to build a close-to-reality model of stress test design by following the literature, such as Parlatore and Philippon (2018) or Parlasca (2019) with multiple banks, each endowed with their own private information and making investment decisions, and the regulator deciding 'how much' information to disclose. Such a quantitative model will be able to guide policymakers on the 'optimal stress test model disclosure' debate.

Chapter 3

Nondisclosure with Conflicting Motives

3.1 Introduction

Why do people have differing beliefs about issues that appear, to an informed audience, to be all but settled with hard evidence? One explanation is that people with access to evidence don't always disclose it, but instead choose to lie by omission to influence others.

Folk wisdom, however, states that hiding information doesn't work if audiences anticipate what's being hidden. Grossman (1981) and Milgrom (1981) provide a classic argument that if the evidence-holder always wants to influence the receiver's action up, then only senders with the lowest signals will consider hiding them, and because of adverse selection, senders can't benefit from silence.

To bridge a gap between this result and the many examples of nondisclosure in the world, this paper explores the possibility that receivers don't know how the sender wants to influence them: they are uncertain about the direction of the sender's preference misalignment. There are a few reasons why this might be the case. First, a receiver may know the identity of the sender, but be uncertain about their preferences on a given issue, either because the issue or the sender herself is unfamiliar. Alternatively, a receiver may not know the sender's identity at all, only that they come from a population of possible informants with different preferences.

My main observation is that, when the receiver doesn't know which direction the sender would like them to bias their action towards relative to their private optimum, then the sender's disclosure policy will never fully unravel. This is because of *bidirectional pooling*: senders with opposite kinds of preferences and evidence on opposite sides of the status quo both end up withholding their information, each relying on the possibility of the other to prevent the receiver from catching on to the nature of the omission. The best the receiver can do in response is take an intermediate action that does well in expectation. In contrast, when the sender's desired direction of influence is known, full disclosure occurs.

Under bidirectional pooling equilibria, the sender discloses some signals, but not others. Their disclosure policies order signal values, and under some additional conditions, the set of sender-preference types, into two-sided spectra with increasingly influential signals and increasingly biased preference types relative to a central signal and central type. Disclosure near the center differs from disclosure at the extremes. Strikingly, two opposite patterns of disclosure arise depending on how much more sensitive the receiver is to the state than the sender. When the receiver and the sender care about the state equally, a sender finds it worthwhile to share highly impactful information, but nudges receivers in the direction of their private bias by omitting details that slightly contradict it. On the other hand, when the receiver's preferred action is significantly more responsive to the state, the sender is reluctant to share big news, because they don't want the receiver to overadjust; instead, they try to influence receivers to take their preferred action by disclosing minor evidence in its favor.

The comparative statics of disclosure from these two cases straddle debates about the importance of heterogeneous perspectives to informative communication. In the first case, greater differences in preferences discourage communication, while in the second, a diversity of perspectives is necessary for unconventional truths to get across.

These results give a framework by which the power to disseminate or withhold facts allows idiosyncratic preferences to affect public outcomes. It is applicable to several important real-world examples. One is media ownership: multiple studies agree that content put forth by media outlets changes with ownership in ways that are consistent with a change in their biased agenda (Gentzkow and Shapiro (2010), Baum and Zhukov (2018)). My results suggest that if media companies are biased relative to the true state of the world but still wish their reports to move with it, then they will all agree on the most essential headlines, but differently biased companies may differ in their coverage by selectively skipping unfavorable minutiae. On the other hand, if they have a state-independent preference for eliciting certain beliefs, then they are reluctant to share any big news, and will instead select their coverage over minor facts that do not generate much "swing".

This paper is laid out as follows. Section 2 gives the main framework of a disclosure model with sender preference uncertainty. Section 3 shows that with enough preference variation, bidirectional pooling equilibria occur. Section 4 discusses the form of the disclosure policy and comparative statics under two main cases. Section 5 works out an example, Section 6 considers the assumptions distinguishing my results from full disclosure, and Section 7 concludes.

3.1.1 Literature review

Models of disclosure typically involve senders who choose to disclose verifiable signals to receivers, who then act upon the information. Classically, in applications such as the quality-signalling problem, imperfect information between the parties lies along a single dimension, that is, the sender's private signal about some shared, payoff-relevant state. Because the sender always wants to influence the receiver's action up, there is adverse selection into

withholding. Any pooling strategies unravel, leading to full disclosure in equilibrium (Milgrom (1981), Grossman (1981)). A general statement of these results, from Okuno-Fujiwara et al. (1990), is that whenever each sender's utility is always strictly increasing (i.e. positive monotone) in each receiver's beliefs about their signal, all state-relevant private information will be revealed.

Strands of the literature have pointed out the possibility of imperfect disclosure when receivers are uncertain about more than the sender's payoff-relevant signal. Dye (1985) observed that, if observers are uncertain about a manager's endowment of information, then those with unfavorable evidence can profitably pool with the uninformed. For the sake of comparison, we could frame the focus of this paper as uncertainty about *what kind* of sender holds useful information, under certainty that the information exists. Banerjee and Somanathan (2001) consider voice in organizations, starting with a model of binary disclosure in which informants may have different priors about the promise of a project. However, in their model, all communication is unidirectional, since the verifiable signal itself is always good news. My paper departs from their binary state/single signal framework and focuses on settings with a continuum of signals and states, and bidirectional communication; however, it shares their focus on the effects of sender heterogeneity.

Also related to the idea of pooling under multidimensional sender heterogeneity are models of costly signaling with privately-known costs (Frankel and Kartik (2019), Esteban and Ray (2006)). There, an informed party with known preferences observes a natural state, as well as their private cost of distorting the decision-maker's perception of the state. Similarly to the sender-specific preferences in this paper, distortion costs are not directly payoff-relevant to the decision-maker, but they confound reports of the state, so that uncertainty about the state remains at the time that decisions are made.

Some more distantly related discussions of cheap talk and partial provability are nevertheless interesting in the context of this problem. Chakraborty and Harbaugh (2010) show that, when there are N dimensions in a message, at most one dimension is payoff-relevant to a sender with fixed preferences, therefore they and the receiver can find an $N-1$ dimensional subspace of common interest on which they can communicate informatively in a cheap talk game. This paper considers something of an opposite case, in which the feasible message is restricted to a single dimension, while the sender's payoff depends on multidimensional information, and shows that this restricts the informativeness of communication. Among models of partial provability, Seidmann and Winter (1997) show that while generically equilibria are partially informative, full revelation occurs when each verifiable subset of types – analogous to a message in my setting – admits a worst-case type, which no other type in that subset wishes to masquerade as. In my setting, sufficient richness of sender preferences and compactness of the signal set precludes the existence of a worst-case type for the empty message.

3.2 Model

I focus on a simple disclosure model with one sender and one receiver, who both aim to maximize their expected utility under utility functions

$$u_s(\theta, x, a_r), \qquad\qquad u_r(\theta, a_r).$$

Payoffs for both players can be directly influenced only through the receiver's action, $a_r \in \mathbb{R}$.

The state of the world, $\theta \in \mathbb{R}$, is unknown, but both players know that its distribution is $f(\theta)$. Conditional on the state, a signal m is drawn at the start of the game from the distribution $h(m|\theta)$.

Assumption 6. *The marginal distribution of the signal, $\int_\theta h(m|\theta) \cdot f(\theta) d\theta$, is continuously supported on a bounded interval $[\underline{m}, \bar{m}]$.*

Finally, the sender has a privately-known preference type $x \sim g(x) \in [\underline{x}, \bar{x}]$, with commonly known distribution independent of $f(\theta)$.[1] For the main body of this paper I assume that $g(x)$ is not a degenerate (point) distribution, but I reconsider this possibility, and its relation to full disclosure results, in the last section.

Assumption 7. *The receiver is uncertain about the sender's type: $g(x)$ is not supported on a single point.*

Though inconsequential in this one-shot model, the underlying idea is that x is intrinsic and known well ahead of time, while m is a signal drawn at the beginning of the game. To reflect this, I will refer to x as simply the sender's *type*, and when referring to a particular sender, I mean a sender endowed with a type x. To avoid confusion, the sender's full set of private information (x, m) (which is what "type" refers to elsewhere in the literature) will instead be called a *scenario* henceforth in this paper.

3.2.1 Timing and actions.

First, the sender's type is drawn, and they learn it. Then, nature draws a state and a signal conditional on it. The sender observes the signal and chooses whether to disclose the signal to the receiver, or to withhold it. I interpret the signal as a piece of hard evidence that can be passed on costlessly to the receiver, and means the same thing to both players. The sender's preference type, on the other hand, is not verifiable, and the sender cannot engage in cheap talk or in any other way influence the receiver's belief about it. Neither the sender nor the receiver has commitment power. Observing only what the sender has passed along, the receiver chooses an action. To summarize, the timing is as follows:

0. x, θ, and m are realized. The sender is given x, and observes m.

[1] Everything will extend to the case where x has unbounded support, as well, but I have chosen to keep x bounded here for ease of exposition.

1. The sender sends a message $\tilde{m}(x, m)$ to the receiver. They may choose between sending their signal as-is ($\tilde{m} = m$), or withholding it ($\tilde{m} = \emptyset$).
2. The receiver observes the message if there was one. They form a Bayesian posterior on the state, which is $\beta(\theta|\emptyset)$ if they saw no message ($\tilde{m} = \emptyset$), and $\tilde{h}(\theta|m)$ if they saw a message $\tilde{m} = m$.
3. The receiver chooses their action $a_r(\tilde{m})$, and payoffs are realized.

3.2.2 Notation and assumptions.

Let $a^*_{r,i}(\cdot)$ denote an optimal choice of a_r from the perspective of player i. That is,

$$a^*_{r,r}(\theta) \in \arg\max \mathbb{E}[u_r(\theta, a)] \qquad a^*_{r,s}(\theta, x) \in \arg\max \mathbb{E}[u_s(\theta, x, a)].$$

I often write a posterior β as an argument in utility function u or maximizer a^*, in place of θ. It is shorthand for taking the expectation over θ given β, e.g.

$$u_r(\beta, a_r) = \mathbb{E}[u_r(\theta, a_r)|\beta] \qquad a^*_{r,r}(\beta) \in \arg\max_a \mathbb{E}[u_r(\theta, a)|\beta].$$

This notation is natural because the state of the world enters the game only through the expectations induced by the signal. Thus, the signal is a "sufficient statistic" with respect to the state and players' strategies, and it is abstractly without loss to take utility functions over realized signals, instead of those over states, as primitives of the model.

In order to impose necessary structure on the basic setup above, I assume that preferences are continuous, differentiable in the action, single-peaked, and ordered in m and x.

Assumption 8. ***Continuity and differentiability:***

$u_s(\theta, x, a_r)$ and $u_r(\theta, a_r)$ are continuous in all arguments, and differentiable in a_r.

Assumption 9. ***Quasiconcavity with increasing peaks (QCIP):***

$u_s(m, x, a_r)$ and $u_r(m, a_r)$ are strictly quasiconcave in a_r, with peaks $a^*_{r,s}(\beta, x)$ strictly increasing in x, and $a^*_{r,r}(\tilde{h}(\theta|m))$ strictly increasing over the family $m \in [\underline{m}, \bar{m}]$.

QCIP means that the utility functions of sender and receiver are both single-peaked, and are ordered with increasing peaks over the possible beliefs induced by signals (for the receiver) and preference types (for the sender), holding the other fixed. Single-peakedness is a common assumption, and increasing peaks is a standard way to order single-peaked functions. The order applies only to the peaks, and preferences are not necessarily well-ordered elsewhere. In particular, this condition does not necessarily imply single crossing differences (SCD), which says that for arbitrary actions $a' > a$, a "higher type" (with higher signal or preference type) will have relatively higher utility for a' rather than a whenever a lower type does.

In fact, Quah and Strulovici (2009) observe that among single-peaked functions, increasing peaks is strictly weaker than SCD, and equivalent to the interval dominance order.

Following the notational discussion above, I assume QCIP directly on the signal-dependent expected utility functions. It can be replaced with an equivalent assumption over the original state-dependent utility functions $u_s(\theta, x, a_r)$ and $u_r(\theta, a_r)$ as long as:

1. $\tilde{h}(\theta|m)$ satisfies the monotone likelihood ratio property, which is sufficient to guarantee a strong set order on the optima.
2. Strict single-peakedness of u_s, u_r is preserved when expectations are taken over $\tilde{h}(\theta|m)$, for all m.

The second condition can often be checked by hand. It is satisfied for a fairly inclusive range of common functional forms. Some useful categories of utility functions and signal structures satisfying (2) are:

- The signal perfectly conveys the state, $m = \theta$.
- Conditional distribution $\tilde{h}(\theta|m)$ is strictly single-peaked for all m, and θ is a shifter of the utility functions, i.e. for some increasing functions γ_s, γ_r,

$$u_s(\theta, x, a + \Delta) = u_s(\theta - \gamma_s(\Delta), x, a), \qquad u_r(\theta, a + \Delta) = u_r(\theta - \gamma_r(\Delta), a).$$

The solution concept I consider is a perfect Bayesian equilibrium (PBE) in pure strategies for the receiver. I will show later that such equilibria always exist here.

PBE is the concept used in most signalling games, including by Grossman and Milgrom for the unraveling result, and in the partial provability literature. In my model, any PBE is pinned down by a single object, which is the receiver's empty-message belief $\beta(\theta|\emptyset)$. The sender best-responds to a given empty-message posterior by choosing between inducing the action the receiver takes upon seeing the true signal, or the one induced by $\beta(\theta|\emptyset)$. Being in a PBE requires that the empty-message posterior be consistent with the state distribution conditional on an empty message, induced by the sender's best response.

Imposing that the receiver play a pure strategy is usually without loss, since under a generic utility function and belief distribution, there will be a single action that maximizes their expected utility. Furthermore, QCIP ensures a single optimal action for the receiver when the signal is revealed to them. However, since my other assumptions will not rule out that there can be a tie for the receiver's expected utility maximizer under the no-message posterior, and my approach relies on the receiver choosing one specific action, in that case, I will force a pure action. This assumption is quite realistic in the direct application to single receivers, since most people don't consciously randomize. It may be less reasonable when the "receiver" stands for the aggregate of a large population, but even then, since at least one pure strategy equilibrium exists, and additional equilibria relying on randomization will be knife's-edge and difficult to sustain, it seems natural to focus on the former.

3.3 Bidirectional pooling

Does the sender always disclose their evidence? In this section, I argue that if the set of possible preferences contain ones that oppose each other under any beliefs for the receiver, then full disclosure never occurs. Formally, the key idea of uncertainty over opposing preferences is a combination of Assumption 6, which establishes type-uncertainty, and Assumption 10 below, which ensures that senders' preferences are sufficiently opposed to rule out full unraveling of nondisclosure.

Assumption 10. ***Bidirectional sender bias (BSB):***

$$\min_m \left[a^*_{r,s}(\beta(\theta|m), \underline{x})\right] < a^*_{r,r}(\underline{m}) \quad \textit{and} \quad a^*_{r,r}(\bar{m}) < \max_m \left[a^*_{r,s}(\beta(m), \bar{x})\right].$$

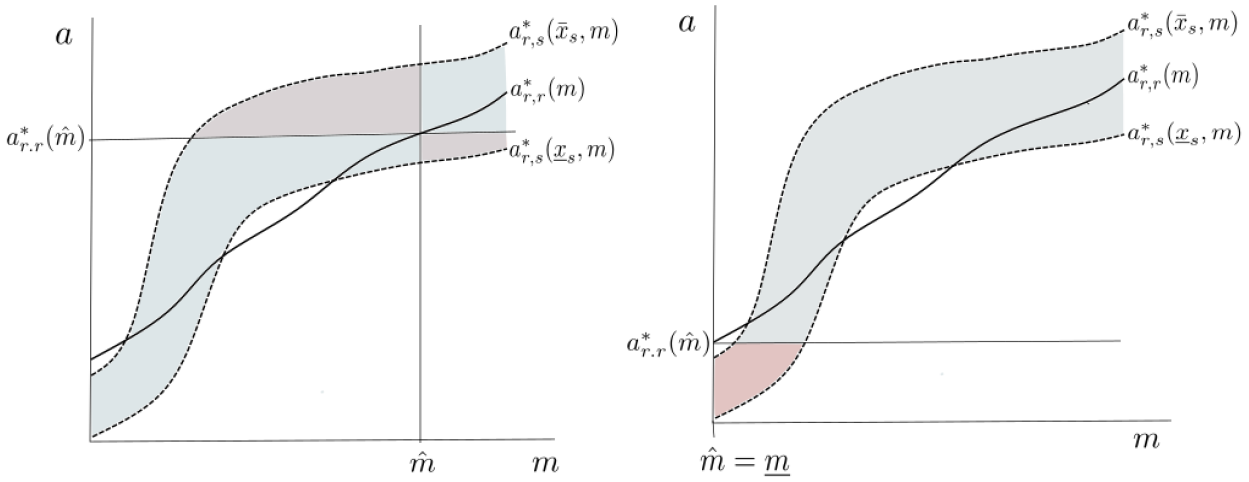
In words, BSB means that the most extreme actions that could be optimal for the sender, over all type and signal realizations, are more extreme than the most extreme optimal actions for the receiver. It captures a strong notion of opposing biases between senders in different scenarios: the receiver is sure that no matter the action they plans to take, there are some scenarios in which the sender wishes it higher, and others in which the sender wishes it lower.

Theorem 10. *Assume that u_s and u_r are continuous in all arguments, differentiable in a_r, and QCIP and BSB are satisfied. Then any signaling equilibrium features $a^*_{r,r}(\beta(\theta|\emptyset)) \in \left(a^*_{r,r}(\tilde{h}(\theta|\underline{m})), a^*_{r,r}(\tilde{h}(\theta|\bar{m}))\right)$, with a positive probability of withholding both "high" signals $(a^*_{r,r}(\beta(\theta|\emptyset)) < a^*_{r,r}(\tilde{h}(\theta|m)))$ and "low" signals $(a^*_{r,r}(\beta(\theta|\emptyset)) > a^*_{r,r}(\tilde{h}(\theta|m)))$. All sender types, except possibly one, withhold under some signal realizations.*

In Figure 3.1, we have **Gray area** $= \{(a^*_{r,s}(x,m), m) : x \in [\underline{x}, \bar{x}], m \in [\underline{m}, \bar{m}]\}$.
Red area $= \{(a^*_{r,s}(x,m), m) : a^*_{r,r}(\hat{m}) \in (a^*_{r,s}(x,m), a^*_{r,r}(m)) \text{ or } (a^*_{r,r}(m), a^*_{r,s}(x,m))\}$.
In equilibrium, $\hat{m} \in (\underline{m}, \bar{m})$, and the sender's best response entails pooling towards $\hat{m}$ from either side. In particular, senders in the red regions will always withhold the signal. Unlike in Grossman and Milgrom, there cannot be a corner posterior given the empty message. If $\hat{m} = \underline{m}$, then the sender would best respond by withholding some higher signals, violating belief consistency.

The only type of sender who may never find it worthwhile to withhold any signal is one who prefers the receiver to take action $a^*_{r,r}(\tilde{\beta}(\theta|\emptyset))$ exactly when their true signal would induce r to take that action anyways.

For a full proof of the theorem, please see the appendix. Here, I will explain the intuition, which is simple. First, fixing the sender's strategy, the receiver's beliefs are also fixed, and their strategy is determined: they play $a^*_{r,r}(\beta(\theta|\tilde{m}))$. There is a signal, $\hat{m} \in [\underline{m}, \bar{m}]$, such that the action taken by the receiver upon seeing the signal $\hat{m}$ is the same as the action taken under $\tilde{m} = \emptyset$: $a^*_{r,r}(\beta(\theta|\emptyset)) = a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$. This signal functions as an endogenously determined "center", the benchmark to which impactful news will be contrasted. It is the status quo not because it represents receivers' prior beliefs, but because it represents the posterior under silence, which can be quite different.



(a) $a^*_{r,s}(\hat{m}, \underline{x}) < a^*_{r,r}(\hat{m}) < a^*_{r,s}(\hat{m}, \bar{x})$ (Pooling)

(b) $a^*_{r,r}(\hat{m}) < a^*_{r,s}(\hat{m}, \underline{x}) < a^*_{r,s}(\hat{m}, \bar{x})$ (Pooling)

Figure 3.1: Outcomes under disclosure and nondisclosure for different configurations of $\hat{m}$.

Recall from our discussion of the model that $\hat{m}$ fixes the equilibrium. To see that any equilibrium will satisfy Theorem 10, observe that under single-peakedness, if $a^*_{r,s}(m, x) > a^*_{r,r}(\hat{m}) > a^*_{r,r}(m)$ or $a^*_{r,s}(m, x) < a^*_{r,r}(\hat{m}) < a^*_{r,r}(m)$, then the sender's strict optimal action is to withhold the signal. It will be helpful to refer to Figure 3.1, where such scenarios appear in red. They are given by the intersection between the 2nd & 4th quadrants of the plane centered on $(\hat{m}, a^*_{r,r}(\hat{m}))$ and the gray area representing the set of possible (signal, sender-optimal action) pairs. Importantly, the assumptions above don't suffice to pin down the *entire* set of scenarios under which nondisclosure is the sender's best move, but the region just described will constitute a strict subset of such scenarios.

It is not hard to see that under BSB, the red region must have positive measure over $g(x) \int_\theta h(m|\theta) f(\theta) d\theta$, and, more importantly, there is a spread over the true value of m across the region. Technical assumption 6 prevents the receiver from taking on beliefs that have probability 0 ex ante and that almost every sender wants to avoid.[2] Thus, senders withhold signals with positive probability, and whenever they do, the receiver is uncertain about the true signal realization.

Single-peakedness also implies that whenever the receiver's belief is consistent and not a singleton, there are both types of senders who prefer to withhold signals above $\hat{m}$ and ones who withhold signals below. Observing that senders' optimal-action curves are functions

[2]If the domain of m were unbounded, and the receiver's belief $b(m|\emptyset)$ allowed to be supported on its closure, then under families of preferences in which the receiver's optimal action varies unboundedly with the message and the sender's utility becomes unboundedly negative with distance from their optimal action, a point belief on $m = \infty$ or $m = -\infty$ would be self-sustaining, due to infinite losses from withholding any finite realization of m. Thus, fully informative equilibria are once again possible. Similar issues arise when m lies in an open interval.

bounded within the gray region of Figure 3.1, it's clear that the only type of sender who may never wish to withhold any signal is the one whose optimal action curve passes through $(\hat{m}, a^*_{r,r}(\hat{m}))$.

Finally, note that while $\hat{m}$ characterizes the equilibrium strategies of sender and receiver uniquely up to indifference, the equilibrium need not be unique, as there may be multiple equilibrium values of $\hat{m}$ that give rise to distinct strategy profiles. An equilibrium in pure strategies for the receiver does always exist, however: a simple intermediate value theorem argument, in conjunction with BSB, shows that a function that takes in $\hat{m}$ and outputs the implied posterior $\hat{m}$ from the sender's BR has a fixed point in $(\underline{m}, \bar{m})$.[3]

3.4 Who withholds information, and when?

The discussion above makes it clear that most types of senders disclose some signals, and withhold others. I now examine which signals each sender withholds, and how that affects the kinds of information that make it through to the receiver. The end result of these comparisons is a set of comparative statics over senders' propensity to communicate and signals' likelihoods of being transmitted, depending on their extremeness relative to special "central" signals and types.

3.4.1 Sender-type monotonicity of disclosure

Are senders more likely to hide impactful information the more it contradicts their bias? Intuition suggests so. By withholding their signal, the sender "corrects" a misalignment between their preferences and the receiver's by letting the receiver carry on with a belief that is slanted relative to the truth, from the sender's point of view. Senders with increasingly extreme low types should be willing to withhold an increasingly large set of signals higher than $\hat{m}$, and senders with increasingly high types should more often withhold signals lower than $\hat{m}$. For the rest of the paper, I will assume single crossing differences, under which this prediction is easy to verify:

Assumption 11. ***Single crossing differences (SCD) in*** x***:*** *For all* m*,* $x < x'$*, and* $a_r < a'_r$*,*

$$\begin{aligned} u_s(\tilde{h}(\theta|m), x, a'_r) - u_s(\tilde{h}(\theta|m), x, a_r) \geq (>)\ 0 \\ \implies u_s(\tilde{h}(\theta|m), x', a'_r) - u_s(\tilde{h}(\theta|m), x', a_r) \geq (>)\ 0. \end{aligned} \tag{3.1}$$

[3] Let $\hat{m}^{BR}(\cdot)$ be an operator taking in a hypothetical value of $\hat{m}$ and outputting the new value of $\hat{m}$ that would represent the receiver's no-message posterior after one round of best responding by the sender. Because the sender's best response is continuous in $\hat{m}$, and the receiver's posterior is continuous in the sender's strategy, $\hat{m}^{BR}$ is continuous in $\hat{m}$. By the argument used to prove Theorem 10, under BSB $\hat{m}^{BR}(\underline{m}) - \underline{m} > 0$ and $\hat{m}^{BR}(\bar{m}) - \bar{m} < 0$; therefore, $\hat{m}^{BR}$ has at least one fixed point in $[\underline{m}, \bar{m}]$.

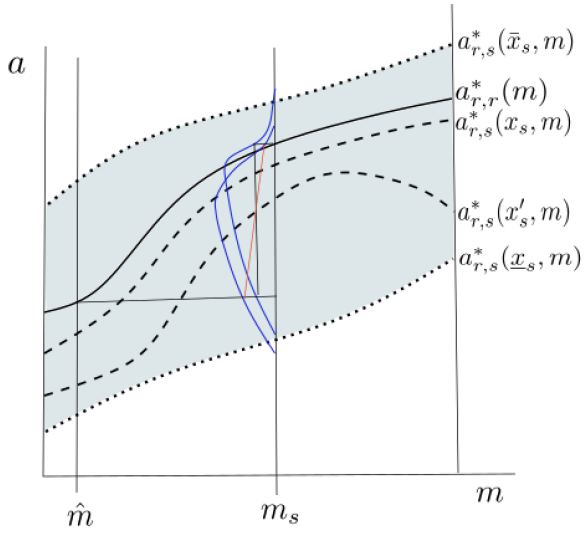


Figure 3.2: SCD guarantees that if a type x is at least indifferent between $\hat{m}$ and $m > \hat{m}$, then a type $x' < x$ will certainly prefer $\hat{m}$, and thus withholds m for sure.

Single crossing differences in x means that if, between a lower action and a higher one, the utility of a sender of lower type is higher for the lower action than for the higher action, then the same is true of the higher type.

Proposition 11. *If, in addition to the assumptions of Theorem 10, u_s satisfies SCD, then the propensity to withhold signals in order to induce a given slant is monotone in sender type: for all m such that $a^*_{r,r}(\tilde{h}(\theta|m)) > (<) a^*_{r,r}(\beta(\theta|\emptyset))$, whenever a sender of type x chooses to withhold m, so do all senders of type $x' < (>) x$.*

Proof. Defining $\hat{m}$ as in the proof of the previous theorem, observe that for all $m < \hat{m}$, SCD in x directly implies that if a type x prefers $\hat{m}$ to m, then a type $x' > x$ does as well, and similarly for $m > \hat{m}$ and $x' < x$. □ □

Proposition 11 states that under single crossing differences, the order on senders' types perfectly captures the (weak) inclusion order on both the set of signals $m < \hat{m}$ that they benefit from withholding, and the set of signals $m > \hat{m}$ that they benefit from disclosing.

Without single crossing differences, there exist counterexamples to this proposition. The reason is that, even if under a given signal m type x' has a preferred action closer to $a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$ and further from $a^*_{r,r}(\tilde{h}(\theta|m))$ than type x, a change in the shape of the rest of the curve may mean that type x' gets *greater* utility than type x from disclosing m, and less from withholding.

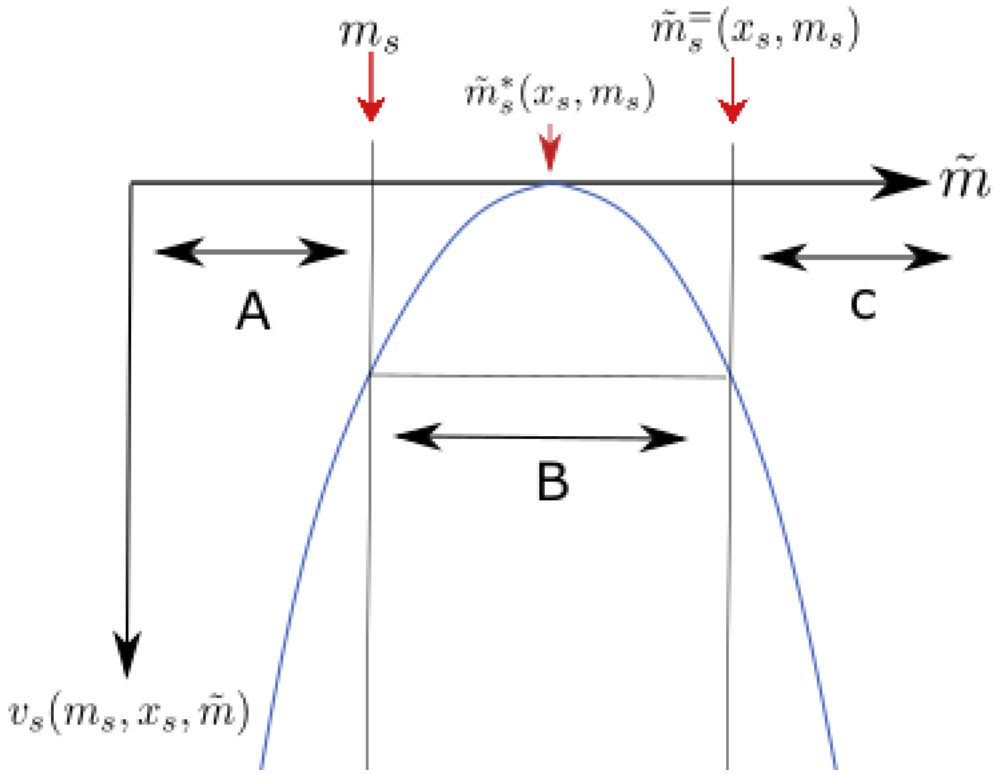


Figure 3.3: Sender's strategy is characterized by nondisclosure in one region, and disclosure in two.

3.4.2 Disclosure policies by type

I now look at disclosure choices within types. Given $\hat{m}$, a full characterization of the disclosure policy for each type is possible.

A few more definitions will be helpful. In particular, since there is a 1-to-1 mapping between disclosed signals and receiver-optimal actions, it will be useful to view the sender's problem as hypothetically maximizing their utility over all possible messages, subject to the disclosure constraint that only $\hat{m}$ and m are actually feasible. Taking as given the receiver's strategy, in the first period the sender chooses a message as if maximizing directly over $\tilde{m}$ the utility function

$$v_s(m, x, \tilde{m}) := u_s(m, x, a^*_{r,r}(\beta(\theta|\tilde{m}))).$$

The function v_s will take on the properties of u_s; in particular, it is single-peaked in $\tilde{m}$. Furthermore, if the sender's choice of message was unrestricted, there would be a unique sender-optimal message

$$\tilde{m}^*(x, m) := \arg\max_{\tilde{m}} v_s(m, x, \tilde{m}).$$

Finally, each sender has a "breakeven message" as a function of true signal m.

Definition 3.4.1. *A* breakeven message $m^{=}(m,x)$ *for the sender is the furthest-away alternative signal that, if sent, would allow the sender to receive at least the same utility as disclosing their true signal:*

$$m^{=}(m,x) = \begin{cases} \min(m \in [\underline{m}, \bar{m}] : v_s(m,x,m') \geq v_s(m,x,m)) & \text{if } m > \tilde{m}^*(x,m) \\ \max(m \in [\underline{m}, \bar{m}] : v_s(m,x,m') \geq v_s(m,x,m)) & \text{if } m < \tilde{m}^*(x,m) \\ m & \text{if } m = \tilde{m}^*(x,m) \end{cases}$$

In Figure 3.3, the value of obfuscating evidence in a given scenario (x,m), when $\hat{m}$ lies in one of 3 regions.

A. Withholding influences beliefs in the wrong direction $\Rightarrow$ disclosure.

B. Profitable nondisclosure.

C. Withholding overcorrects in the direction of bias $\Rightarrow$ disclosure.

Single-peakedness of $v_s(m,x,\cdot)$ implies that an alternative hypothetical message is preferred to m if and only if it lies between m and $m^{=}(m,x)$. Figure 3.3 shows why: when $\hat{m}$ lies towards $\tilde{m}^*(x,m)$ relative to m, nondisclosure directionally favors the sender's bias, but if it is to the other side of $\tilde{m}^{=}(x,m)$, then the omission goes too far.

Therefore, fixing $\hat{m}$, the sender's strategy, up to indifference at the boundaries, is

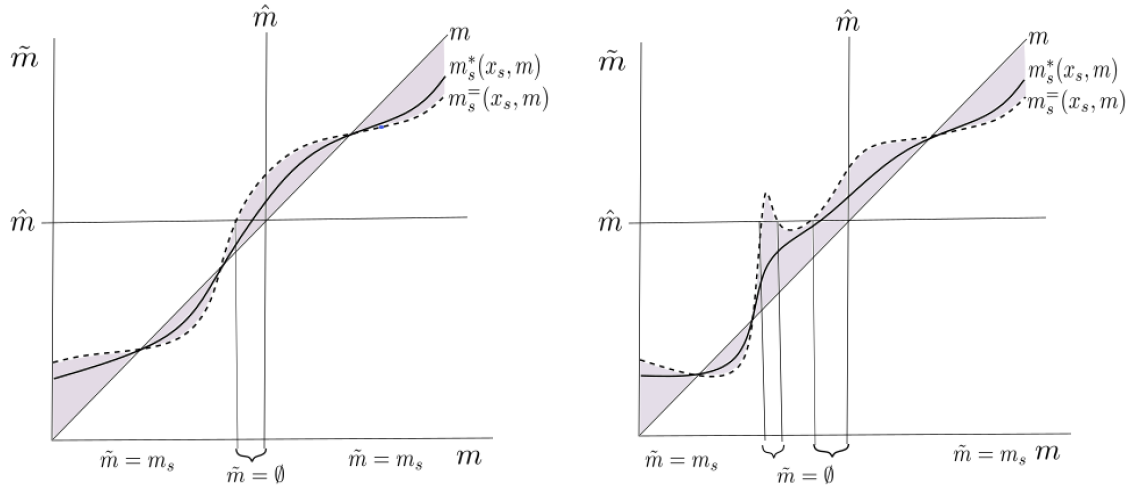
$$\tilde{m}(m,x) = \begin{cases} \emptyset & \text{if } m \leq m^{=}(m,x) \text{ and } \hat{m} \in [m, m^{=}(m,x)] \\ & \text{or } m \geq m^{=}(m,x) \text{ and } \hat{m} \in [m^{=}(m,x), m] \\ m & \text{otherwise.} \end{cases} \tag{3.2}$$

Figures 3.4a and 3.4b give examples of this concept. Fixing a sender, the intersection of $\tilde{m} = \hat{m}$ with the purple region between the true and breakeven messages gives the set of signals the sender will withhold.

3.4.3 Monotone breakeven message

When $m^{=}$ is well-behaved, we can make fairly straightforward predictions about what kinds of signals senders disclose. The multi-segmented policy of Figure 3.4b occurs because the breakeven message in that case is nonmonotone, which can occur when the shape of the sender's utility function or the amount of misalignment with the receiver is highly state-dependent. On the other hand, when the breakeven message is monotone, strategies involving a single interval of nondisclosure will be the only possible outcome.

Proposition 12 (Montone breakeven message)**.** *In addition to previous assumptions, if for all* x*,* $\tilde{m}^{=}(m,x)$ *is strictly increasing in* m*, then*



(a) Monotonicity of the breakeven message implies an interval withholding strategy.

(b) An optimal strategy for the sender when the breakeven message is not monotone in m.

Figure 3.4: Breakeven messages plotted against true signals.

- *There is $\hat{x}$ such that the sender of type $\hat{x}$ is indifferent between disclosing or withholding $\hat{m}$ and discloses everything else.*
- *As x increases from $\hat{x}$, senders withhold an increasing interval of signals $[m^*, \hat{m}]$, and as x decreases from $\hat{x}$, senders withhold an increasing interval of signals $[\hat{m}, m^*]$.*

On the other hand, if for all x, $\tilde{m}^=(m, x)$ is strictly decreasing in m, then

- *There is $\hat{x}$ such that the sender of type $\hat{x}$ is indifferent between disclosing or withholding $\hat{m}$ and withholds everything else.*
- *As x increases from $\hat{x}$, senders disclose an increasing interval of signals $[m^*, \hat{m}]$, and as x decreases from $\hat{x}$, senders disclose an increasing interval of signals $[\hat{m}, m^*]$.*

Figure 3.4a illustrates the concept. Intuitively, when $m^=(\cdot, x)$ is monotone increasing, the threshold for overshooting the sender's bias is increasing in the realized signal, while the reverse holds when it is monotone decreasing. Loosely speaking, the sender's and receiver's interests are relatively aligned when $m^+(\cdot, x)$ is increasing, whereas they are misaligned when it is decreasing. Neither implies, nor is implied by single crossing differences or QCIP.

Both the positive and negative monotone cases fit some simple cases. Positive monotonicity in $m^+(\cdot, x)$ tends to occur when senders are misaligned due to simple bias. Formally, I define a simple bias setting to be one in which the sender's utility function is simply shifted relative to the receiver's by a bias function ω increasing in x:

$$u_s(\theta, x, a_r) = u_r(\theta, a_r - \omega(x)).$$

The breakeven message will always be positive monotone when preference misalignment takes the form of simple bias and:

1. u_s is symmetric about $a^*_{r,s}(m, x)$, or
2. m shifts u_s and u_r together: there is a single increasing function γ such that

$$u_s(m, x, a + \Delta) = u_s(m - \gamma(\Delta), x, a), \qquad u_r(m, a + \Delta) = u_r(m - \gamma(\Delta), a).$$

Bias, symmetry, and signals as shifters are all common in models of policy targeting and principal-agent models of delegation with communication.

On the other hand, an important class of problems for which the breakeven message is negative monotone is that in which the sender's preferences are completely state-independent. That is, senders could be completely dogmatic, or the state could be payoff-relevant only to receivers, even though the sender cares about the realized outcome. Examples include lobbyists and interest groups, or strict ideologues.

Section 6 gives an example of monotone increasing $m^+(\cdot, x)$ in a situation with pure bias, symmetry, and a state-matching motive, as well as an example with negative MBM when the sender is very insensitive to the state, relative to the receiver.

3.4.4 Comparative statics under monotonicity

Having established settings in which positive- and negative-monotone in $m^+(\cdot, x)$ are reasonable assumptions, I turn to highlight some comparative statics of the probability of disclosure in these cases. Observe that signals and types are very much bidirectional, with the "center" of each bidirectional spectrum naturally defined by $\hat{m}$ and $\hat{x}$, respectively. Symmetry across the center means that the most interesting comparative statics will be about *extremeness*, or distance from the center, rather than about high vs. low signals or types. In what follows, I assume without much loss that indifferent senders always choose disclosure.[4]

The first set of comparative statics concern the communicativeness of senders. There is exactly one central type $\hat{x}$, who can be considered a pure centrist. Under positive montonicity, this type's interests are aligned enough with the receiver's to want to disclose everything (even though their preferences and the receiver's generally differ). With a negative montonicity, the centrist's interests are not particularly aligned with the receiver's, nor do they have a particular interest in championing a cause; thus, they never disclose anything. As a corollary of Prop. 11, the total probability of disclosure is monotone with distance from x to either side:

Corollary. *When $m^+(\cdot, x)$ is monotone increasing, the sender's total probability of disclosing a signal is quasiconcave in x and maximized at $\hat{x}$.*

[4]This does not change the set of equilibria, nor their properties

When $m^+(\cdot, x)$ is monotone decreasing, the sender's total probability of disclosing a signal is quasiconvex in x and minimized at $\hat{x}$.

So, in cases where senders' and receivers' misalignment approximates a simple bias, extreme senders are, on the whole, less likely to disclose a message. Fixing receivers' beliefs, a greater spread in the distribution of sender types decreases the amount of communication. On the other hand, when senders' preferences are less state-sensitive than the audience's, only extreme types are willing to share influential information, and more dispersed preferences lead to a greater flow of information.

The contrast between these two cases relates to a debate about the positive or negative impacts of diverse values. A common thought is that polarization can jam communication. Many issues for the receiver, such as decreased trust or lack of common ground, contribute, but the positive-monotone example supports the idea that the extreme sender's reluctance to communicate also plays a role. The pattern under the negative-monotone example, on the other hand, echoes an argument in favor of pluralism made by Banerjee and Somanathan (2001).[5] There, a greater variety in types is associated with more communication, because it takes an agreeably biased sender to support the transmission of any major news – strong partisan bases are necessary to bring truths to light.

It matters not just that different senders can be more or less forthcoming, but also that some signals may be disclosed more often than others. Abstracting away from the realization of senders' types, differential transmissibility of signals directly determines welfare and outcomes on the receiver side. A corollary of Proposition 12 is that:

Corollary. *When $m^+(\cdot, x)$ is monotone increasing, a signal's total probability of being disclosed is quasiconvex in m except at $\hat{m}$, where it is 1. As $m \uparrow \hat{m}$ or $m \downarrow \hat{m}$, the probability of disclosure decreases.*

When $m^+(\cdot, x)$ is monotone decreasing, a signal's total probability of being disclosed is quasiconcave in m and maximized at $\hat{m}$.

More extreme signals are more transmissible under positive monotonicity because senders disclose all bias-favoring signals. An interpretation of this is that bigger news is more likely to travel because everyone agrees on the importance of publicizing it, whereas biased sources are all too happy to sway audiences by fudging the small stuff. The contrasting prediction under negative monotonicity is that extreme signals are less likely to be transmitted through disclosure. They are too influential, and must be withheld for fear of causing the receiver to overreact.

The preceding discussion of conditions for positive and negative monotone breakeven messages suggests that one distinguishing factor is the sender's state-sensitivity relative to

[5]Banerjee and Somanathan's verifiable communication model actually satisfies conditions for a *positive* monotone breakeven message, but their signals are unidirectional by design, and for them, an extreme sender is one who has a private desire to support a project, which corresponds in my model to a higher, not a more misaligned, type.

the receiver. With a specific setting in mind, this distinction can help settle the debate both about the transmissibility of extreme signals and about extreme senders' communicativeness.

3.5 Example: policy platforms.

Following the discussion of equilibrium policies, I provide an illustrative example. Suppose that a receiver follows a member of the press (sender) on Twitter, and would, in an upcoming election, like to support one of a continuum of government spending policies, which range from very austere (-1) to very expansionary (1). There is a factor, $\theta \sim U[-1,1]$, that influences the optimal level of government spending, but it is unknown to both agents. The sender is affected by what the receiver does through its impact on the outcome of the election. In addition, each agent may be in a position to benefit privately from government programs, or to suffer from higher taxes, so I model their utilities as a quadratic loss function

$$u_s(\theta, x, a_r) = -(a_r - \theta - x)^2, \qquad u_r(\theta, a_r) = -(a_r - c\theta)^2.$$

with the sender's private preference parameter x uniform on $[-1,1]$. Finally, suppose that the sender has access to briefing with information relevant to θ that can be summarized as a signal $m \sim U[\theta - \epsilon, \theta + \epsilon]$, and may disclose it verifiably.

The utility functions in this example satisfy continuity, differentiability, and BSB; the distributions satisfy Assumptions 6 and 7, and both jointly satisfy QCIP and SCD. Thus, there will be bidirectional pooling and an ordering of disclosure in sender types. When $c < 2$, the breakeven message exhibits positive monotonicity, and when $c > 2$, it exhibits negative monotonicity.

I solve for the full disclosure policy to illustrate the outstanding characteristics of each case, starting with a description of the best response of the receiver and sender when the other side's strategy is fixed.

Receiver's decision: $a^*_{r,r}(\tilde{m}) = c\mu(\theta|\tilde{m})$, where the mean posterior message is either, if $\tilde{m} = m$,

$$\mu(\theta|m) = \begin{cases} m, & m \in [-1+\epsilon, 1-\epsilon] \\ \frac{1+m-\epsilon}{2}, & m \in [1-\epsilon, 1+\epsilon] \\ \frac{-1+m+\epsilon}{2}, & m \in [-1-\epsilon, -1+\epsilon] \end{cases}$$

or, if the message is empty, $\mu(\theta|\emptyset) = \int_{\underline{x}}^{\bar{x}} \int_{\underline{m}}^{\bar{m}} 1_{\tilde{m}(x,m)=\emptyset} \mu(\theta|m) dm dx$.

Sender's decision: Fix the mean posterior upon seeing nothing, $\mu(\theta|\emptyset)$. Then

$$\tilde{m}(x,m) = \begin{cases} \emptyset & \text{if } x > 0 \text{ and } \mu(\theta|\emptyset) - \mu(\theta|m) \in [0, \frac{2x}{c}] \\ & \quad \text{or } x < 0 \text{ and } \mu(\theta|\emptyset) - \mu(\theta|m) \in [\frac{2x}{c}, 0] \\ m & \text{else.} \end{cases}$$

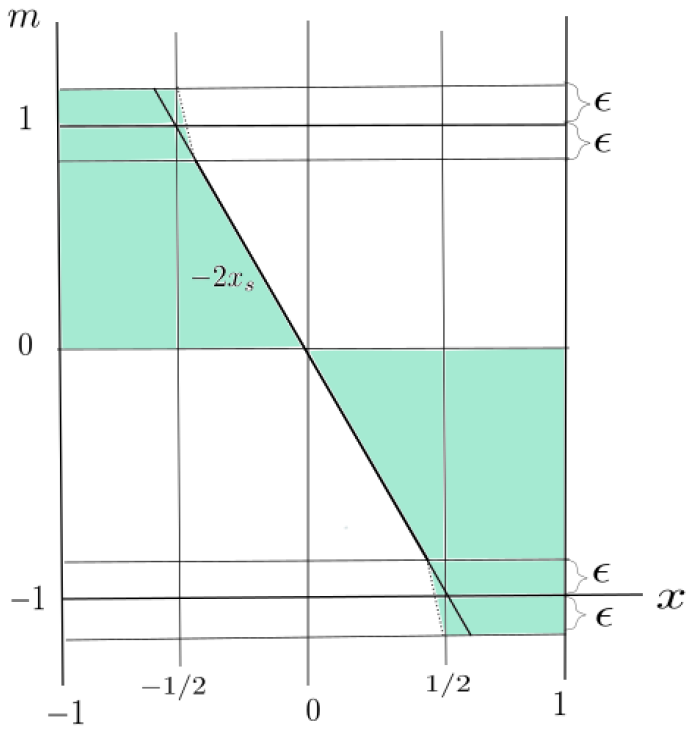


Figure 3.5: $c = 1$

The belief $\mu(\theta|\emptyset)$ is determined in equilibrium, and fully characterizes the equilibrium actions through the above best responses.

Claim 10. *For all c, the unique equilibrium consistent with a Bayesian receiver is given by the above actions and $\mu(\theta|\emptyset) = 0$.*

Figure 3.5 shows when the sender chooses to disclose or withhold signals when $c = 1$. This is a "pure bias" setting in which senders and receivers are equally sensitive to the state, but the sender's preferences are offset from the receiver's based on private preferences. Withholding occurs in the shaded area.

In contrast, when $c > 1$ the sender is more moderate relative to the receiver, introducing a new dimension of misalignment between the players that varies with the magnitude of the signal m. The case $c = 2$ is liminal, and disclosure policies are particularly simple here: senders disclose everything aligned with their bias, and hide everything opposed. When the sender is much less sensitive to the state than the receiver, i.e. $c > 2$, we approach the case where, relative to the receiver, the sender is almost state-agnostic: extreme signals are rarely revealed, and only the most extreme senders are willing to disclose them.

We can contrast this outcome with those possible when the receiver is certain about how the sender wishes to influence them:

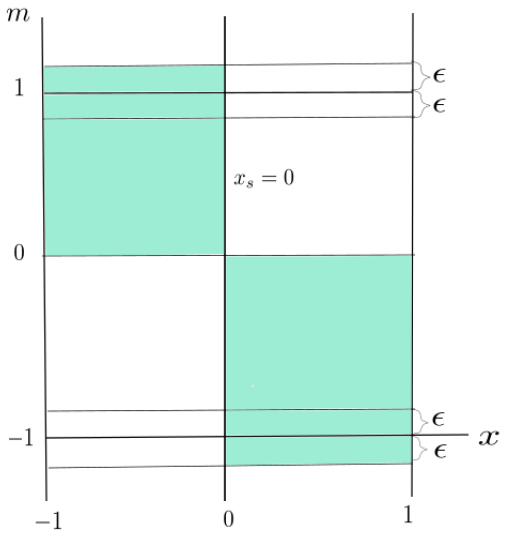


Figure 3.6: $c = 2$

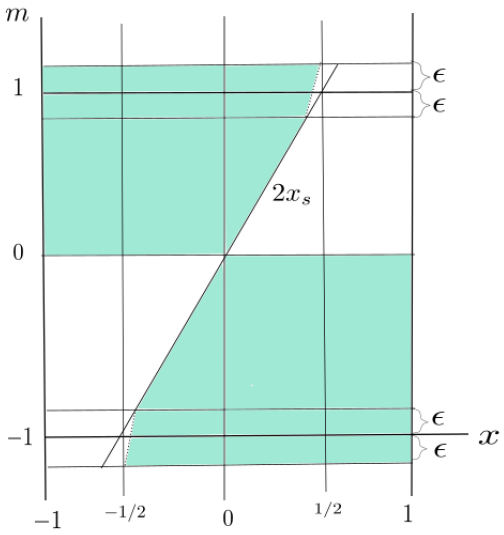


Figure 3.7: $c = 3$

Claim 11. *If, in this example, the sign of the sender's preference type x is known, then in the unique equilibrium when $0 < c < 2$, the sender's messages are fully separated according to the realized signal.*

Knowledge of sender type alone does not guarantee full disclosure, and in particular, Claim 11 is not true if $c > 2$: as Claim 13 in the following section will show, monotone-increasing $m^{=}(\cdot, x)$ sufficient to guarantee full disclosure under sender-type certainty, but it turns out monotone-decreasing $m^{=}(\cdot, x)$ is not sufficient for either full disclosure or full nondisclosure.[6]

3.6 Full separation with unidirectionality or certainty

The main way in which my model departs from the literature is by assuming that senders' preferences are both uncertain and bidirectional. Plenty of models assume the opposite, and obtain full disclosure. A well-known result allowing some variation in sender preferences by type is the monotonicity theorem of Okuno-Fujiwara et al. (1990), which states that full disclosure is the unique possible outcome whenever senders' utilities are monotone in the receiver's beliefs over the entire space of scenarios. Though intuitive, this theorem is not a particularly good fit to settings in which senders' preferences have a single peak in the interior of the action space. More applicable is the idea from Seidmann and Winter (1997) that full disclosure is a possible outcome if and only if each possible message admits a different "worst-case" scenario, which senders in no other scenario would like to pretend to be. Their logic applied here shows that if BSB fails dramatically, in that the sender-optimal

[6]Partial disclosure strategies are possible under negative monotonicity.

action is either always greater or always less than the receiver-optimal action in any scenario, then full disclosure is the unique outcome.

Claim 12. *If for all m, x, $a^*_{r,r}(\tilde{h}(m)) < (a^*_{r,s}(\tilde{h}(m), x))$, then $\hat{m} = \underline{m}$, and in the unique equilibrium the sender reveals fully reveals their signal by choosing $\tilde{m} = m$ under all signal realizations. A similar argument holds when $a^*_{r,r}(\tilde{h}(m)) > \max_x(a^*_{r,s}(\tilde{h}(m), x))$.*

What happens if, instead, the sender's bias could go either way, but is deterministic given their signal, with the mapping from signal to preferred action known by the receiver? Because misalignment between the sender's and receiver's preferences depends on the state, this in and of itself does not guarantee full disclosure. Nevertheless, when $m^+(\cdot, x)$ is monotone-increasing, all equilibria are fully revealing (and such an equilibrium exists).

Claim 13. *If x is known to the receiver, and the breakeven message is positive monotone, then an equilibrium exists, and in any equilibrium m is fully revealed.*

3.7 Conclusion

Can a sender with unknown objectives and access to hard evidence influence others through their choice to disclose or withhold evidence? The answer depends on how the audience updates their beliefs under nondisclosure. This paper shows that when a sender could potentially have either of two opposing biases, receivers can't fully back out the sender's evidence or their identity. Therefore, relative to a symmetric information benchmark, ownership of evidence benefits the sender by allowing them to withhold some unfavorable news. In cases where senders and receivers are similarly sensitive to the state, but senders have a state-independent bias, strong signals tend to be revealed, whereas weak ones are often hidden, especially by heavily biased sources. Alternatively, when senders are agnostic to the state, they avoid disclosing strong signals unless also strongly biased.

A couple of extensions of this model are straightforward. I have considered preference uncertainty, and Dye (1985) considers imperfect disclosure under uncertainty about information endowments. If there is uncertainty about *both*, then imperfect disclosure will still occur, and nondisclosing senders will pool with both informed senders under opposite scenarios, and the uninformed. The set of equilibria will differ from that without uncertainty in informedness, but the form of equilibrium and comparative statics will follow what I have outlined in this paper: the intuition is that information endowment uncertainty changes the posterior under silence, but it doesn't change informed senders' best responses conditional on the center.

Another potential addition is receiver-preference uncertainty. Among other things, allowing receivers' preferences to vary is natural in an anonymous setting where neither the sender nor the receiver's identity is known. Passing articles over the internet is a nice example of this. When receivers' preferences aggregate into a utility function satisfying the single-receiver conditions, my conclusions carry over immediately. In some work omitted

here, I show that when receiver types are well-ordered and each type's expected utility depends on the state only through its expectation, pooling is also guaranteed under similar conditions.

Directions for future work include evaluating the impact of certain assumptions. I have assumed no cheap talk about the sender's preferences, but in practice communication about preferences may sometimes be possible, and may alter disclosure. In addition, signals may be divisible, allowing the sender some freedom in the degree of disclosure beyond a binary choice. I've also assumed that senders' preference types are payoff-irrelevant to the receiver, but they could instead be thought of as a payoff-relevant signal that cannot be disclosed in a game of multidimensional communication. Finally, although many instances of disclosure, such as those in a courtroom, or regarding the viability of a short-term opportunity, are well approximated by a one-shot game, some relationships between informants and audiences are long-lived. In these repeated games, persistent uncertainty about preferences may be less sustainable, but it would be interesting to explore the possibility that senders may still exercise some power by building a reputation.

Appendix A

for "Inference from selectively disclosed data"

A.1 Construction of the imitation equilibrium

We will prove that Theorem 2 holds in a more general case with potentially state-contingent, rather than state-independent, data-mass distributions. Describe a game in this general setting by $\mathcal{G}(\Theta, \mathcal{D}, \beta_0, \{f_j\}_{j=1}^J, d\{G^j\}_{j=1}^J)$ where G^j describes the distribution of μ under state j. The model we describe in the main text corresponds to the case in which $G^j = G$ for all j.

Theorem 13. *Suppose that $g^1, \ldots, g^J$, the densities of μ under states $\theta_1, \ldots, \theta_J$, respectively, are continuous on $\mathbb{R}$ and supported on $[0, 1]$. There exists a unique imitation equilibrium outcome, implemented by a vector-valued burden of proof function $\hat{\boldsymbol{\mu}}(u) : [0, \theta_J] \to \mathbb{R}^J$ with inverse $\hat{u}_k(\mu)$ such that*

1. *$\hat{u}_j(\mu)$ is continuous and (weakly) increasing in μ for all j.*
2. *$\sigma^*(\mu f_j)$ is supported on $\{\mu' f_k : \mu' = \hat{\mu}_k(\hat{u}_k(\mu / r_j(k)))$ and $\theta_k \in A_j(\mu)\}$.*

To outline the argument, we first prove the existence of a imitation equilibrium by construction. Then we prove the separation theorem, which we use to show uniqueness.

Recall that $\hat{u}_k(\mu)$ is the equilibrium payoff to sending the message μf_k.

We construct $\hat{u}_k(\mu)$ that is monotone increasing in μ – 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 $\hat{u}_k$, which we denote by $\hat{u}_k^-$ 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 $v = \theta_J$.

Recall that $r_j(k) = \max_{d\in\mathcal{D}} \frac{f_k(d)}{f_j(d)}$ is the ratio of the amount of data necessary to imitate a certain amount of f_k under state j to the amount necessary under state k, and

$$A_j(\mu) = \left\{\theta_k : k \in \arg\max_{k>j} \hat{u}_k\left(\frac{\mu}{r_j(k)}\right)\right\}.$$

is the set of states that type μf_j finds it weakly optimal to target given $\hat{\boldsymbol{\mu}}$.

The range of $\hat{u}_k(\mu_k)$ 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 under which the receiver optimally takes an action that yields the sender 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 $\hat{\boldsymbol{\mu}}(v) = (\hat{\mu}_k(v))_{\theta_k\in S(v)}$, the associated *frontier* consists of all types that are just able to meet some component of $\hat{\boldsymbol{\mu}}(v)$ with no slack, that is, all types $\tilde{\mu}_j f_j$ such that

$$r_j(k)\tilde{\mu}_j = \hat{\mu}_k(v) \text{ for some } \theta_k \in S(v), \text{ and } \nexists \theta_{k'} \in S(v) \text{ s.t. } r_j(k')\tilde{\mu}_j > \hat{\mu}_{k'}(v). \tag{A.1}$$

Given a particular burden of proof function $\hat{\boldsymbol{\mu}}$, the implied frontier for payoff v is $\tilde{\mu}(v|\hat{\boldsymbol{\mu}}) = (\tilde{\mu}_1, \ldots, \tilde{\mu}_{l-1}, \hat{\mu}_l(v), \ldots, \hat{\mu}_J(v))$ if $S(v) = \{\theta_l, \ldots, \hat{\theta}_J\}$ where $\tilde{\mu}_1, \ldots, \tilde{\mu}_{l-1}$ satisfy eq. **??**.

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

$$\tau^{opt}_{\hat{\boldsymbol{\mu}}}(\theta_k, v) = \{\theta_j : f_k \in A_j(\tilde{\mu}_j(v|\hat{\boldsymbol{\mu}}))$$

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

$$\tau^{supp}_{\hat{\boldsymbol{\mu}}}(\theta_k, v) = \{\theta_j : \hat{\mu}_k(v)f_k \in \text{supp } \sigma(\cdot|\mu f_j) \text{ for some } \mu\}.$$

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

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, $\tau^{opt}_{\hat{\boldsymbol{\mu}}}(S, v) = \bigcup_{\theta_k\in S} \tau^{opt}_{\hat{\boldsymbol{\mu}}}(\theta_k, v)$ and $\tau^{supp}_{\hat{\boldsymbol{\mu}}}(S, v) = \bigcup_{\theta_k\in S} \tau^{supp}_{\hat{\boldsymbol{\mu}}}(\theta_k, v)$, and for every set $\omega \subseteq [0, 1]$, we let $A_j(\omega) = \bigcup_{\mu\in\omega} A_j(\mu)$.

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

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

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

$$W_{\hat{\boldsymbol{\mu}}}(S, v|\sigma) = \frac{\sum_{\theta_j \in \tau^{opt}_{\hat{\boldsymbol{\mu}}}(S,v)} \beta_0(\theta_j)\theta_j g^j(\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]) \frac{d\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]}{dv} \sigma(\{\hat{\mu}_k f_k\}_{\theta_k \in S} | \tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)] f_j)}{\sum_{\theta_j \in \tau^{opt}_{\hat{\boldsymbol{\mu}}}(S,v)} \beta_0(\theta_j) g^j(\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]) \frac{d\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]}{dv} \sigma(\{\hat{\mu}_k f_k\}_{\theta_k \in S} | \tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)] f_j)} = v. \quad \text{(A.2)}$$

For any partial strategy $\hat{\sigma}$ that gives mixing probabilities between the messages $\hat{\mu}_{k_i}(v)\mathbf{f}$, the payoff $W_{\hat{\boldsymbol{\mu}}}(S, v|\hat{\sigma}(v))$ is always weakly greater than $V_{\hat{\boldsymbol{\mu}}}(S, v)$. 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 $\tau^{opt}_{\hat{\boldsymbol{\mu}}}(\theta_k, v) \bigcap \tau^{opt}_{\hat{\boldsymbol{\mu}}}(\theta_{k'}, v) \neq \emptyset$, 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 $\hat{\boldsymbol{\mu}}(v)$.

Algorithm: This algorithm calculates the payoffs to targeting a state in $S(v)$ at frontier $\hat{\boldsymbol{\mu}}(v)$ 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 $\theta_k, \theta_{k'} \in \Sigma$,

$$\frac{r_j(k)}{r_j(k')} = \frac{\hat{\mu}_k(u)}{\hat{\mu}_{k'}(u)} = \frac{\frac{d\hat{\mu}_k(u)}{du}}{\frac{d\hat{\mu}_{k'}(u)}{du}} \quad \left(= \frac{\frac{d\hat{u}_{k'}(\hat{\mu}_{k'}(u))}{d\mu}}{\frac{d\hat{u}_k(\hat{\mu}_k(u))}{d\mu}} \right) \quad \text{(A.3)}$$

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

We define

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

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 $\hat{\boldsymbol{\mu}}(v)$, that have an amount $\hat{\alpha}\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]$ of data, when we assume that eq. A.3 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 $\tau^{supp}_{\hat{\boldsymbol{\mu}}}(\Sigma, v)$, 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 A.3 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 types $\alpha\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]f_j$ for some $\theta_j \in \tau^{supp}_{\hat{\boldsymbol{\mu}}}(v)$ pooled with each other, then payoffs are equal to

$$\hat{u}_k(\alpha\hat{\mu}_k(v)) = v_\Sigma(\alpha, \hat{\boldsymbol{\mu}}(v)) \equiv \frac{\sum_{\theta_j\in\tau^{supp}_{\hat{\boldsymbol{\mu}}}(\Sigma,v)}\beta_0(\theta_j)\theta_j g\left(\alpha\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]\right)\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]}{\sum_{\theta_j\in\tau^{supp}_{\hat{\boldsymbol{\mu}}}(\Sigma,v)}\beta_0(\theta_j) g\left(\alpha\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]\right)\tilde{\mu}_j[\hat{\boldsymbol{\mu}}(v)]}\Bigg|_{\alpha=\hat{\alpha}},$$

which is continuous in α because g is continuous everywhere in $(0,1]$.[1] The burden-of-proof function for $\underline{v} \leq v$ is then given by

$$\mu^{put}_\Sigma(\underline{v}) \equiv \{v^{-1}_\Sigma(\underline{v}, \hat{\boldsymbol{\mu}}(v))\hat{\mu}_k(v)f_k\}_{\theta_k\in S(v)},$$

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

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

[1] It is important that $g(1) = 0$, since this ensures that g^j is continuous at $\mu = 1$.

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 $V_{\hat{\boldsymbol{\mu}}}(\Sigma', \hat{\boldsymbol{\mu}}(v)) < V_{\hat{\boldsymbol{\mu}}}(\hat{\Sigma}, \hat{\boldsymbol{\mu}}(v))$ 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 $V_{\hat{\boldsymbol{\mu}}}(\hat{\Sigma}, \hat{\boldsymbol{\mu}}(v))$, there must be some message that yields payoff at least $V_{\hat{\boldsymbol{\mu}}}(\Sigma, \hat{\boldsymbol{\mu}}(v))$. But since there is no message, and indeed no proper subset of messages in $\hat{\Sigma}$ that achieve payoff $V_{\hat{\boldsymbol{\mu}}}(\Sigma, \hat{\boldsymbol{\mu}}(v))$ 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 $\hat{\boldsymbol{\mu}}(v)$ 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 $\hat{\boldsymbol{\mu}}(v)$, 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 $\hat{\boldsymbol{\mu}}(v)$.

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

Claim 14. *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 $\hat{\boldsymbol{\mu}}(v)$, 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' $v_{\Sigma'}^{-1}(v'', \hat{\boldsymbol{\mu}}(v))\hat{\mu}_k(v)f_k \tilde{\subseteq} t$ for some $v'' > v'$*
2. *There is a partition element Σ with a subset $\Sigma' \subseteq \Sigma$ such that $v_{\Sigma'}(v_\Sigma^{-1}(v', \hat{\boldsymbol{\mu}}(v)), \hat{\boldsymbol{\mu}}(v)) > v'$,*

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

Note that, because for any partition element $\Sigma' \neq \Sigma$ either $v_{\Sigma'}^{-1}(v, \hat{\boldsymbol{\mu}}(v))\hat{\mu}_k(v)f_k \tilde{\not\subseteq} t$, 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 $v_\Sigma(\alpha, \hat{\boldsymbol{\mu}}(v))$ 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 $\Sigma' \subseteq \Sigma$, $v_{\Sigma'}(v_\Sigma^{-1}(v, \hat{\boldsymbol{\mu}}(v)), \hat{\boldsymbol{\mu}}(v)) \leq v$ 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. For every reasonable example we can think of, the number of such points (and thus steps in the construction) is not just countable, but 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 $\hat{u}_k(\mu_k)$ is increasing, all types capable of sending a message in $\{\hat{\mu}_j(v)\mathbf{f}_j\}_{j=1}^J$ achieve a payoff of at least v. We define the set of types that are incapable of sending a message in $\{\hat{\mu}_j(v)\mathbf{f}_j\}_{j=1}^J$, 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,

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

Note that $\sup_M U(T(v, M)) \geq v$, because $\lim_{\alpha\to 1} U(T(v, \alpha\hat{\boldsymbol{\mu}})) = v$. 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 $\hat{T}^{max}_{\hat{\boldsymbol{\mu}}}(v)$. 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 $\hat{T}^{max}_{\hat{\boldsymbol{\mu}}}(v)$ is well-defined.
2. Whenever $\hat{T}^{max}_{\hat{\boldsymbol{\mu}}}(v)$ exists, there exist $\mu_l, \ldots, \mu_J$ such that $\hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v) = T(v, \{\mu_l f_l, \ldots, \mu_J f_J\})$.
3. Whenever $\hat{T}^{max}_{\hat{\boldsymbol{\mu}}}(v)$ exists, there exists a partial strategy $\hat{\sigma} : \hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v) \to M = \{\mu_l f_l, \ldots, \mu_J f_J\}$ such that the payoff to any message $m \in M$ given that senders in $\hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v)$ play according to $\hat{\sigma}$ is $\hat{U}_{\hat{\boldsymbol{\mu}}}(v)$.

The first point follows from the fact that, unless the union of two such sets yields payoff at least $\hat{U}_{\hat{\boldsymbol{\mu}}}(v)$, 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 $\hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v)$ is a maximum-payoff pool, then for each subset $S \subseteq M$, the

payoff to the pool implemented by S is no greater than $U(\hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v))$, which is sufficient to ensure that $\hat{\sigma}$ exists. In addition, $U(T(v, M))$ is absolutely continuous with respect to every component of $\hat{\boldsymbol{\mu}}(v)$ and each μ_k.

Lemma 14. *If $\hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v)$ exists, then Condition 1 is satisfied by the burden of proof vector $M = \{\mu_l f_l, \ldots, \mu_J f_J\}$ such that $\hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v) = T(v, M)$.*

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

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

   Then, since $V_{\hat{\boldsymbol{\mu}}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J)$ is continuous in α, there is $\underline{\alpha} < 1$ such that $V_{\hat{\boldsymbol{\mu}}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J) > v$ 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 $(\alpha M)_{\alpha=\underline{\alpha}}^1$ 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 $T(M, v) = \hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v)$.

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

   Then WLOG let Σ' be the lowest-value element of the partition. Similarly to the above, since $V_{\hat{\boldsymbol{\mu}}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J)$ is continuous in α, there is $\bar{\alpha} > 1$ such that $V_{\hat{\boldsymbol{\mu}}}(\Sigma, \alpha(\mu_k f_k)_{k=l}^J) < v$ 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 $T(M, v) = \hat{T}^{\max}_{\hat{\boldsymbol{\mu}}}(v)$.

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 2 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 $\hat{\boldsymbol{\mu}}(\theta_l)$, as constructed from the previous step. For all $v > \theta_l$, let $\hat{\boldsymbol{\mu}}(v)$ be as already constructed. Define

$$\check{\mu}_l(\theta_l) = \max\{\mu : \exists j < l \text{ s.t. } \tilde{\mu}_j[\hat{\boldsymbol{\mu}}(\theta_l)] \geq \mu\},$$

and rewrite $\hat{\boldsymbol{\mu}}(\theta_l) = (\check{\mu}_l(\theta_l), \hat{\mu}_{l+1}(\theta_l), \ldots, \hat{\mu}_J(\theta_l))$. Proceed as below to rewrite $\hat{\boldsymbol{\mu}}(v)$ for $v < \theta_l$:

(b) Fix $S = \{\theta_k\}_{k=l}^J$. Given the frontier $\hat{\boldsymbol{\mu}}(v)$, check if $\hat{T}^{max}_{\hat{\boldsymbol{\mu}}}(v)$ exists, and if so, find $M = \{\mu_l f_l, \ldots, \mu_J f_J\}$ that implements $\hat{T}^{max}_{\hat{\boldsymbol{\mu}}}(v)$ and rewrite $\hat{\boldsymbol{\mu}}(v) = M$.

(c) At $\hat{\boldsymbol{\mu}}(v)$, using the algorithm, partition S into subsets of states, and calculate $v_\Sigma(\alpha, \hat{\boldsymbol{\mu}}(v))$ for all $\alpha \in [0,1]$ for each subset. Take the lowest-value frontier, $\hat{\boldsymbol{\mu}}(v')$, under putative payoffs $v_\Sigma(\alpha, \hat{\boldsymbol{\mu}}(v))$ such that the conditions of Claim 14 are satisfied and such that $\hat{T}^{max}_{\hat{\boldsymbol{\mu}}}(v'')$ does not exist for any $v'' \in (v', v]$, and assign strategies according to Algorithm 2 between $\hat{\boldsymbol{\mu}}(v)$ and the new frontier $\hat{\boldsymbol{\mu}}(v')$.

(d) Set $v = v'$ and set $\hat{\boldsymbol{\mu}}(v')$ 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 $\hat{\boldsymbol{\mu}}$.

The existence of an imitation equilibrium, and the monotonicity of $\hat{u}_k$, follow directly from this construction. Continuity of $\hat{u}_k$ also follows from this construction. The value of $\hat{u}_k$ is defined on series of closed intervals on each of which it is continuous – $v_\Sigma(\alpha, \mu)$ is continuous in α, and $\hat{u}_k(\mu)$ is constant for $\mu f_k \in T(v, M)$. Together, these cover the domain of u_k, that is, $[0,1]$, and they overlap only at their endpoints, at which they coincide.

A.2 Properties of σ^*

Here, we prove a collection of results about the structure of the imitation equilibrium. We begin by showing the separation theorem holds, which we use to show that u_{σ^*} is unique. We then proceed to give proofs of other results in sections 1.2 and 1.4.

A.2.1 Proof of separation theorem and uniqueness

First, we prove the separation theorem. It has 2 parts, which we will prove as lemmas. We start by proving that upper pools are improving:

Lemma 15. *If M is a collection of messages and $\{\tilde{\mu}_j(v)f_j\}_{j=1}^J$ is the frontier of types achieving a payoff of at least v under σ^*, where $\theta_i < v \leq \theta_{i+1}$, then*

$$\mathbb{E}_q[\theta | t \in U(\{\underline{\mu}_j f_j\}_{j=1}^J) \setminus U(M)] \geq v$$

whenever $U(\{\tilde{\mu}_j f_j\}_{j=1}^J) \setminus U(M)$ is nonempty.

Proof of Lemma. Denote $T(v, M) = U(\{\tilde{\mu}_j f_j\}_{j=1}^J) \setminus U(M)$. Let $(\bar{\mu}_1, \ldots, \bar{\mu}_i; \bar{\mu}_{i+1}, \ldots, \bar{\mu}_J)$ be the minimum masses of data distributed like $f_1, \ldots, f_i; f_{i+1}, \ldots, f_J$, respectively, necessary to send some message in M. Then

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

If $(\bar{\mu}_{i+1}, \ldots, \bar{\mu}_J) \leq (\tilde{\mu}_{i+1}, \ldots, \tilde{\mu}_J)$ pointwise, then $T(v, M)$ is empty. Otherwise, let the states $j_1, \ldots, j_A$ be the maximal set such that $(\bar{\mu}_{j_1}, \ldots, \bar{\mu}_{j_A}) > (\tilde{\mu}_{j_1}, \ldots, \tilde{\mu}_{j_A})$ pointwise. Call the set of types that send $\mu' f_{j_a}$ with positive probability under σ^* by $\tau_{\sigma^*}^{supp}(\mu' f_{j_a})$, and let $\theta(t)$ refer to the state corresponding to the distribution of dataset t. Denote by $\hat{\sigma}_v$ the partial strategy, restricting to types in $T(v, M)$, where those types play as they do in σ^*, and assume that the receiver knows the sender is in $T(v, M)$ and playing according to this strategy.

Let ϕ_{σ^*} be a joint density over types and messages induced by σ^*, so that for type $t = \mu f_j$ and message $m = \tilde{\mu} f_{\tilde{j}_a}$, we can define

$$\phi(t, m) = g^j(\mu r_j(\tilde{j}))\sigma^*(m|t)\beta_0(\theta_j)r_{\theta_j}(j')$$

to be the density on the event that the sender is type t and plays message m, when t plays m with positive probability. In the case when payoffs under u_{σ^*} are strictly increasing at $\mu' f_{j_a}$, each sender who plays $\mu' f_{j_a}$ is randomizing between at most a finite number of messages in their mixed strategy, one corresponding to each state that is weakly optimal for them to imitate. Thus, they play each message in the support of their strategy with strictly positive probability, rather than randomizing with some density over a continuum of messages; ϕ therefore fully captures the distribution of play for senders playing $\mu' f_{j_a}$.

When payoffs are strictly increasing at $\mu' f_{j_a}$, we know that for every $\mu' f_{j_a}$ that is in $T(v, M)$ and is on-path in σ^*, the receiver's inference when they know the sender's type is in $T(v, M)$ in addition to knowing they played message $\mu' f_{j_a}$ is weakly better than if they only know $\mu' f_{j_a}$ was the message played. Formally,

$$\begin{aligned}\mathbb{E}_q[\theta|\mu' f_{j_a}] &= \frac{\sum_{t\in\tau_{\sigma^*}^{supp}(\mu' f_{j_a})\bigcap T(v,M)} \theta(t)\phi(t,\mu' f_{j_a}))}{\sum_{t\in\tau_{\sigma^*}^{supp}(\mu' f_{j_a})\bigcap T(v,M)} \phi(t,\mu' f_{j_a})} \\ &\geq \frac{\sum_{t\in\tau_{\sigma^*}^{supp}(\mu' f_{j_a})} \theta(t)\phi(t,\mu' f_{j_a})}{\sum_{t\in\tau_{\sigma^*}^{supp}(\mu' f_{j_a}))} \phi(t,\mu' f_{j_a})} \\ &\geq v\end{aligned} \tag{A.4}$$

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

Since, of course, payoffs under u_{σ^*} may not be strictly increasing at every $\mu' f_{j_a}$ in $T(v, M)$, we have to separately consider the case in which they are constant, i.e. the case where there are positive-measure pools T of senders achieving the same payoff $v' > v$ under σ^* with $T \bigcap T(v, M)$ nonempty. Then let M' be the set of messages that implements the pool, and

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

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

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. □ □

Next we prove that the imitation equilibrium we construct has worsening lower pools. This is relatively simple.

Lemma 16. *If M is a collection of messages and $\{\tilde{\mu}_j(v) f_j\}_{j=1}^J$ is the frontier of types achieving a payoff of at least v under σ^*, where $\theta_i < v \leq \theta_{i+1}$, then*

$$\mathbb{E}_q[\theta|t \in U(M) \setminus U(\{\tilde{\mu}_j f_j\}_{j=1}^J)] < v$$

whenever $U(M) \setminus U(\{\tilde{\mu}_j f_j\}_{j=1}^J)$ is nonempty.

Proof. If there was a payoff frontier $\hat{\boldsymbol{\mu}}(v)$ that had a nonempty, weakly improving lower pool lower-bounded by messages M, then there is a frontier $\hat{\boldsymbol{\mu}}(w) \neq M$ for some $w \geq v$ such that

$$u_{pool}(U(M) \setminus U(\hat{\boldsymbol{\mu}}(w))) = w.$$

The construction algorithm rules this out, because if indeed the payoff frontiers above $\hat{\boldsymbol{\mu}}(w)$ are correctly constructed, then it would next set $\hat{\boldsymbol{\mu}}(w) = M$. □

Finally, we show that the constructed equilibrium outcome is the only imitation equilibrium outcome, and thus that the imitation equilibrium outcome is unique.

Proof. Let the constructed equilibrium be σ^*, and let σ be an alternative equilibrium, with a different outcome. We aim to show that σ^* does not have improving upper pools, and therefore cannot be an imitation equilibrium.

To see this, let M represent the frontier of messages that are used to achieve payoff v in σ. Worsening lower pools under σ^* imply that $u_{pool}(U(M) \setminus U(\hat{\boldsymbol{\mu}}(v))) \leq v$, implying that M has a worsening upper pool. Since M is a payoff frontier of σ, the alternative equilibrium σ does not have improving upper pools, and is therefore not an imitation equilibrium. □

A.2.2 Proof of results in Section 1.2

Next, we formally show that the imitation equilibrium satisfies the 3 selection criteria that we discuss in Section 1.2: credible inclusive announcement-proofness, truth-leaning, and receiver-optimality.

First we discuss a way in which the imitation equilibrium outcome arises from optimal behavior for the sender. The concept of optimality we use, *inclusive* announcement-proofness, refines PBE by requiring that there is no self-separating set of sender types who could weakly improve their payoffs by announcing a strategy that uses some set of messages differently than they are used in the baseline equilibrium.

Definition A.2.1. *Given an outcome u_{σ^*}, a set of types T has a* credible inclusive announcement *that they will play a strategy $\hat{\sigma}_M$ supported 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 $\mathbb{E}_{\beta_{\hat{\sigma}_M}(\cdot|m)}[\theta] = v$ for all $m \in M$.*
- *$T = \{t \in U(M) : u_{\sigma^*}(t)\}$, and there is some $t \in T$ with $u_\sigma(t) < v$.*

A very closely-related notion, that we take the name from, is the idea of a credible announcement, from Matthews et al. (1991). There is, however, a subtle difference, which is that in a credible announcement, $T = \{t \in U(M) : u_{\sigma^*}(t) \leq v\} \bigcup S$ where $S \subseteq \{t \in U(M) : u_{\sigma^*}(t) = v\}$. Thus what we use is an "inclusive" notion of a credible announcement in that the set of announcing types must include all who weakly prefer to participate; it is stronger to claim there exists an credible inclusive announcement than that there exists a credible announcement, and correspondingly, inclusive announcement-proofness is weaker than announcement-proofness. In fact, there may exist no announcement-proof equilibrium

at all in the game we study, while there always exists exactly one inclusive announcement-proof equilibrium outcome.

Claim 15. *In $\mathcal{G}$, the unique inclusive announcement-proof equilibrium outcome is the imitation equilibrium outcome.*

Proof. For any equilibrium σ with a different outcome than the imitation-equilibrium outcome σ^*, there is some v such that the v-payoff frontier under σ differs from that under σ^*, and such that some types that achieve a payoff of v or greater under σ^* achieve a payoff no more than v under σ. Lemma 15 ensures that when all such types pool, the expected value of the state is at least v. Then, from the continuity of $\hat{u}_j(\mu)$, there exists some $v' < v$ such that when the set of all types that achieve a payoff of at least v' under σ^*, but a payoff of no more than v under σ, is pooled, the expected value of the state is exactly v. Starting from equilibrium σ, this set of types has a credible inclusive announcement that yields a payoff of v to each type, and so σ is not inclusive announcement-proof.

On the other hand, any credible announcement relative to baseline equilibrium σ^* requires the existence of some v and set of messages M such that there exists a pool of types

$$T = \{t \in U(M) : u_{\sigma^*}(t) \leq v\}$$

such that $\mathbb{E}[\theta | t \in T] = v$, with at least one type $t' \in T$ such that $u_{\sigma^*}(t') < v$. Since T contains all types $t \tilde{\supset} t'$ with $u_{\sigma^*}(t) \leq v$, we know T is a set of positive measure. The construction algorithm for σ^*, however, rules out the presence of any such set T, since if all frontiers for payoffs in $(v, \theta_J]$ are correctly constructed, then all types in T must be pooled under σ^* and must obtain a payoff of v exactly. □

We prove that truth-leaning equilibria and imitation equilibria coincide in $\mathcal{G}$, that the imitation equilibrium outcome is unique, and that it is the optimal outcome of communication under commitment for the receiver.

Claim 16. *Every imitation equilibrium of $\mathcal{G}$ is a truth-leaning equilibrium of $\mathcal{G}$.*

Proof. We take the 2 perturbations separately. First, perturb the likelihood of honest commitment types by a sequence with $\epsilon^k_{t|t} = \epsilon^k \to 0$. There exists an equilibrium $u_{\sigma^*_{\epsilon^k}}$ of $\mathcal{G}^{\epsilon^k}$ in which strategies of non-commitment types are identical to the imitation equilibrium strategies in a game $\tilde{\mathcal{G}}^{\epsilon^k}$ under which

$$q(\mu f_j) = \begin{cases} \frac{\beta_0(\theta_j)(g(\mu) - \epsilon^k)}{1 - \epsilon^k \sum_i \beta_i (1 - G^i(\hat{\mu}_i(\theta_i)))}, & \mu \geq \hat{\mu}_j(\theta_j) \\ \frac{\beta_0(\theta_i) g(\mu)}{1 - \epsilon^k \sum_i \beta_i (1 - G^i(\hat{\mu}_j(\theta_i)))}, & \mu < \hat{\mu}_j(\theta_j). \end{cases}$$

Under the metric induced by the L2 norm, the set of equilibrium strategies is compact, and payoffs in $\tilde{\mathcal{G}}^\epsilon$ are continuous in ϵ, so the limit point as $k \to \infty$ of the imitation equilibria of $\tilde{\mathcal{G}}^{\epsilon^k}$ must also be an equilibrium of $\mathcal{G}$. It is easy to verify that it must also satisfy the conditions in 1.2.1, so it is the imitation equilibrium of $\mathcal{G}$.

Now, for fixed ϵ_k, consider in addition the perturbation of payoffs by an additional payoff bump ν to a truthful report. When $\nu < \min_{j,k} |\theta_k - \theta_j|$, there exists an equilibrium $\sigma^*_{\epsilon^k,\nu}$ that is identical to the equilibrium $u_{\sigma^*_{\epsilon^k}}$ specified above, except for types μf_j with $u_{\sigma^*_{\epsilon^k}} \in (\theta_j, \theta_j + \nu)$, who instead play the truth with positive probability. In particular, for a given message $\mu' f_k$ that yields a payoff in $(\theta_j, \theta_j + \nu)$ and is played by μf_j under $\sigma^*_{\epsilon^k}$, the probability that it is played by μf_j in the equilibrium of the further-perturbed game is 0 if the expected state over types playing $\mu' f_k$ for whom the state is not θ_j is no greater than $\theta_j + \nu$, and otherwise, the probability that μf_j plays μf_k is exactly such that the payoff to playing μf_k is $\theta_j + \nu$, so that μf_j is indifferent between playing message $\mu' f_k$ and revealing all their data. As $\nu \to 0$, the set of affected types shrinks towards a measure-0 set, and so these equilibria converge to $u_{\sigma^*_{\epsilon^k}}$ as $\nu \to 0$.

Finally, given the equilibria $\{\sigma^*_{\epsilon^k,\nu^j}\}$ for $\epsilon^k \to 0$, $\nu^j \to 0$, diagonalize by taking, for every k, some j_k such that $||\sigma^*_{\epsilon^k,\nu^{j_k}} - \sigma^*_{\epsilon^k}|| < \frac{1}{k}$, and observe that then the sequence of perturbations $(\epsilon_{t|t} = \epsilon^k \forall t, \epsilon_t = \nu^{j_k} \forall t)_{k=1}^{\infty}$ yields equilibria that converge to σ^*. □

Claim 17. *Every truth-leaning equilibrium in $\mathcal{G}$ is an imitation equilibrium of $\mathcal{G}$.*

Proof. If t's dataset is off-path then the receiver plays a best response to the belief $\mathbb{1}_t$ upon seeing t. This suffices to show that every truth-leaning equilibrium messaging strategy σ is a best response to q_σ, as defined by eq. A.5:

$$q_{\sigma^*}(t|m) := \begin{cases} \frac{q(t)\sigma^*(m|t)}{\sum_{t \in \mathcal{T}_N} q(t)\sigma^*(m|t)} & \text{for on-path } m, \\ \mathbb{1}_m & \text{for off-path } m \in \mathcal{T}_N \\ q_{\sigma^*}(t | \underset{t' \supseteq m,\ t' \in \mathcal{T}_N}{\arg\min} \ \mathbb{E}_{\beta(\cdot|\sigma^*(t'))}[\theta]) & \text{for off-path } m \in \mathcal{M}_N \setminus \mathcal{T}_N. \end{cases} \tag{A.5}$$

For part a), note that if a message m is on-path in σ, then there exists K_1 such that for all $k > K_1$, m is on-path in $\sigma^*_{\epsilon_k}$. For every k, however, all on-path messages are in $\mathcal{T}$, since if m is on-path and $m \notin \mathcal{T}$, then there is a type $t = \mu f_j$ with $\theta_j > u_{\sigma^*_{\epsilon_k}}(m)$ that plays m, and t itself is not played as a message on path by any non-commitment types. But then $\mathbb{E}_{\beta_{\sigma^*_{\epsilon^k}}(\cdot|t)}[\theta] = \mathbb{E}_{\pi(\cdot|t)}[\theta] \geq \mathbb{E}_{\beta_{\sigma^*_{\epsilon_k}}(\cdot|m)}[\theta]$, leading to a contradiction. Hence, all on-path m must be in $\mathcal{T}$.

To prove that a truth-leaning equilibrium messaging strategy satisfies c), suppose there is t such that $\mathbb{E}_{\pi(\cdot|t)}[\theta] > \max_{m \tilde{\subseteq} t} \mathbb{E}_{\beta_\sigma(\cdot|m)}[\theta]$ but $\sigma(t|t) < 1$.

. We will show that there is no sequence of perturbations $\{\epsilon^k_t, \epsilon^k_{t|t}\}_{k=1}^{\infty} \to 0$ such that equilibria of the associated perturbed games $\mathcal{G}^k$ converge to σ. Start by supposing for the sake of contradiction that there is. First, we know t must be on path in σ. If σ^k is an equilibrium of game $\mathcal{G}^k$ with $\epsilon^k_t > 0$, there cannot $t' \neq t$ such that $\sigma^k(t|t') > 0$, otherwise

$\mathbb{E}_{\beta_{\sigma^k}(\cdot|t)}[\theta] \geq \max_{t' \tilde{\subset} t} \mathbb{E}_{\beta_{\sigma^k}(\cdot|t')}[\theta]$ and so $\mathbb{E}_{\beta_{\sigma^k}(\cdot|t)}[\theta] + \epsilon_t^k > \max_{t' \tilde{\subset} t} \mathbb{E}_{\beta_{\sigma^k}(\cdot|t')}[\theta]$ and we would have to have $\sigma^k(t|t) = 1$. Then, likewise, in the limit σ, we must have $\sigma(t|t') = 0$ for all t'. Since t is on-path in σ, it must be that $\sigma(t|t) \in (0, 1)$.

Take a type $t'' \neq t$ such that $\sigma(t''|t) > 0$. We know that there exists K such that for all $k > K$, $\sigma^k(t''|t) > 0$ as well. Then whenever $k > K$, $\mathbb{E}_{\pi(\cdot|t)} + \epsilon_t^k = \mathbb{E}_{\beta_{\sigma^k}(\cdot|t'')}[\theta]$. Because $\sigma^k \to \sigma$, we have that

$$\lim_{k \to \infty} \mathbb{E}_{\beta_{\sigma^k}(\cdot|t'')}[\theta] = \mathbb{E}_{\beta_{\sigma}(\cdot|t'')}[\theta] = \max_{m \tilde{\subseteq} t} \mathbb{E}_{\beta_{\sigma}(\cdot|m)}[\theta].$$

But this contradicts that $\mathbb{E}_{\pi(\cdot|t)}[\theta] > \max_{m \tilde{\subseteq} t} \mathbb{E}_{\beta_{\sigma}(\cdot|m)}[\theta]$ and

$$\lim_{k \to \infty} \mathbb{E}_{\beta_{\sigma^k}(\cdot|t'')}[\theta] = \lim_{k \to \infty} \mathbb{E}_{\pi(\cdot|t)} + \epsilon_t^k = \mathbb{E}_{\pi(\cdot|t)}.$$

To show that b) holds, note that for any k, if t is on-path and played by some $t' \neq t$, then $\sigma^k(t|t) = 1$. By c), $\mathbb{E}_{\pi(\cdot|t')} \leq \mathbb{E}_{\beta_{\sigma^k}(\cdot|t)}$, but if t also plays t and $\mathbb{E}_{\pi(\cdot|t)} < \mathbb{E}_{\beta_{\sigma^k}(\cdot|t)}$, then the receiver cannot Bayesian. On the other hand, if t is on-path and only t plays t, then we must have $\mathbb{E}_{\pi(\cdot|t)} = \mathbb{E}_{\beta_{\sigma^k}(\cdot|t)}$. □ □

Finally, closely following the idea in Hart et al. (2017), we show that the imitation equilibrium outcome is the outcome of the optimal pure-strategy mechanism, that is, the best outcome the receiver can achieve when they can commit to a pure action as a response to the message the sender sends. The revelation principle shows that it suffices to look at direct mechanisms, in which the sender truthfully reports their type and the receiver commits to a deterministic response to the sender's reported type.

A mechanism under which type t elicits the action $a(t)$ is implementable if it satisfies IC:

$$t \tilde{\subseteq} t' \;\Rightarrow\; a(t') \geq a(t). \tag{IC}$$

Claim 18. *The imitation equilibrium outcome is the optimal outcome for the receiver under commitment to pure strategies.*

To prove this claim, first define $T_{\mu f_k}$ be the set of types that imitate μf_k under σ^*, including μf_k itself. We start with a lemma.

Lemma 17. *There always exists an imitation equilibrium σ^* such that $T_{\mu f_k}$ is finite for every $\mu f_k \in \mathcal{T}$.*

Proof of Lemma 17. First, for any imitation equilibrium, if $\{t : u_{\sigma^*}(t) = u_{\sigma^*}(\mu f_k)\}$ is a measure-0 set, since then it is necessarily true that at most one type under each state lies in the same payoff frontier as μf_k under σ^*, and thus at most one type under each state imitates it.

Now consider the case in which there is a positive-measure set of senders who achieve the payoff $u^* = u_{\sigma^*}(\mu f_k)$, where we have $\theta_l \leq u_{\sigma^*}(\mu f_k) < \theta_{l+1}$. We know that there exists a way to divide the types by which state they imitate, and with what probability, given by sets $S_{l+1}, \ldots, S_J$ and any imitation equilibrium σ^*, such that

$$\frac{\sum_{t \in S_j} \theta(t) q(t) \int_{\hat{\mu}_j(u^*)}^{\inf_{v>u^*} \hat{\mu}_j(v)} \sigma^*(\mu f_j | t) d\mu}{\sum_{t \in S_j} q(t) \int_{\hat{\mu}_j(u^*)}^{\inf_{v>u^*} \hat{\mu}_j(v)} \sigma^*(\mu f_j | t) d\mu} = u^*$$

and for all $\mu^* \in (\hat{\mu}_j(u^*), \inf_{v>u^*} \hat{\mu}_j(v))$,

$$\frac{\sum_{t \in S_j : t \tilde{\supseteq} \mu^*} \theta(t) q(t) \int_{\hat{\mu}_j(u^*)}^{\inf_{v>u^*} \hat{\mu}_j(v)} \sigma^*(\mu f_j | t) d\mu}{\sum_{t \in S_j : t \tilde{\supset} \mu^*} q(t) \int_{\hat{\mu}_j(u^*)}^{\inf_{v>u^*} \hat{\mu}_j(v)} \sigma^*(\mu f_j | t) d\mu} \leq u^*.$$

But it is always feasible to reorder the imitation strategy to construct σ^{**} such that $S_{l+1}, \ldots, S_J$ are unchanged, but if $\mu_1 f_j$ imitates $\mu_1' f_i$ and $\mu_2 f_j$ imitates $\mu_2' f_i$, with $\mu_1 > \mu_2$, then $\mu_1' > \mu_2'$ also. That is, conditional on imitating the same state, higher-data senders always imitate types with more data under σ^{**}. Then any type is imitated by either a single type or an interval of types under any other state; the latter is ruled out by the fact that it would result in a payoff no more than θ_l to the message. Once again, since there is a finite set of states, this ensures that each type is imitated by at most a finite set of other types. □

Proof of Claim 18. Suppose A to be the subset of types in $\mathcal{T}$ that are imitated under the imitation equilibrium σ^*, and suppose that σ^* is an imitation equilibrium in which each type is imitated by a finite set of other types, which exists by the previous lemma. Given $\mu f_j \in A$, let $T_{\mu f_j}$ be the set of types that play μf_j under σ^*, including μf_j itself. Define a distribution over $T_{\mu f_j}$,

$$q_{\mu f_j}(t) = \frac{q(t)\sigma(\mu f_j | t)}{\sum_{t \in T_{\mu f_j}} q(t)\sigma(\mu f_j | t)},$$

which is the probability of type t conditional on the message μf_j.

Call the optimal direct mechanism a^*, that responds with the action $a^*(t)$ after receiving the report t. It must satisfy IC across any subset of types, $T \subseteq \mathcal{T}$, but let us consider instead w, the solution to a relaxed local problem where we impose that IC must hold only between $t, t' \in T_{\mu f_j}$ when types are distributed according to $q_{\mu f_j}$. We will show that for all $t \in T_{\mu f_j}$, we have $w(t) = \mathbb{E}_{q_{\mu f_j}}[\theta]$, and that taking this solution across all $\mu f_j \in A$ assigns a response for the receiver to all $t \in T$ while preserving global IC, and therefore gives the optimal direct mechanism.

We know that $w(\mu f_j) \leq w(t)$ for all $t \in T_{\mu f_j}$. Let $S_{\mu f_j} = \{t \in T_{\mu f_j} : w(t) = w(\mu f_j)\}$. First, note that if $S_{\mu f_j} = T_{\mu f_j}$, then we optimally have $w(t) = \mathbb{E}_{q_{\mu f_j}}[\theta]$ for all $t \in T$. This leaves us to rule out that $w(t) \neq w(t')$ for some $t, t' \in T_{\mu f_j}$.

We rule out that $w(\mu f_j) \geq \mathbb{E}_{q_{\mu f_j}}[\theta]$ and $w(t) \neq w(\mu f_j)$ for some $t \in T_{\mu f_j}$, due to the fact that the receiver can then improve their payoff while preserving IC by instead responding to every type with $w(\mu f_j)$. Next, we rule out that $w(\mu f_j) < \mathbb{E}_{q_{\mu f_j}}[\theta]$ and $w(t) \neq v(\mu f_j)$ for some $t \in T_{\mu f_j}$, since then it is possible to instead respond to every t such that $w(t) = w(\mu f_j)$ with $\min_{t \in T_{\mu f_j} \setminus S} w(t)$, and, by single-peakedness of the receiver's payoff function, this improves the receiver's payoff.

This suffices to show that w corresponds exactly to the outcome of the imitation equilibrium for all $t \in T_{\mu f_j}$, regardless of the choice of $\mu f_j \in A$. As w optimizes the receiver's payoff under a weaker set of IC constraints than a^*, we know that the imitation equilibrium outcome is at least as good as a^* for the receiver; the reverse statement is immediate since every equilibrium outcome is implementable with commitment, and so the two are identical. □ □

Corollary (to Claim 18)**.** *The imitation equilibrium outcome is the receiver-optimal equilibrium outcome.*

Proof. In every equilibrium σ, the receiver has a unique best response to each message, given by the action

$$a_r(\beta(\cdot|m))) = \mathbb{E}_{\beta(\cdot|m)}[\theta].$$

Any type of the sender therefore has an optimal feasible message to send that results in a unique optimal action that they can induce the receiver to take given the receiver's inference function. Any equilibrium outcome can therefore be implemented by the receiver through a direct mechanism that responds to every type with a deterministic message, and so there is no equilibrium that increases the receiver's payoff relative to the optimal pure-strategy mechanism outcome that is also the imitation equilibrium outcome. □ □

A.2.3 Proof of results in section 1.4

Convergence to full-information outcome as $Var(g) \to 0$

Proof of Claim 2. We show that given any infinite sequence of games with data-mass distributions $g_1, g_2, \ldots$ on $[0,1]$ with a fixed mean and variances $Var_1, Var_2, \ldots \to 0$, that are identical in the set of states and their ex-ante distribution, the payoff to a sender conditional on the state converges in probability to their full-information payoff.

In order to do so, we show that for any δ and ϵ, there exists L such that for all $l \geq L$, the distribution g_l is such that $Pr[u_{\sigma^*}(\mu, \theta_k) < \theta_k - \delta] < \epsilon$ under every state.

Define the mean of μ to be $\bar{\mu}$, and

$$B = \max_{j \neq k} \frac{1}{r_j(k)}$$

so that for any two states j and k, the difference between the amount of the state-k distribution that the mean type under state k has and the amount the mean type under state j has is $\bar{\mu}(1-B)$.

Suppose that the variance of μ under density g_L is less than $\Delta^2\epsilon^2$, where $\Delta > 0$ is an arbitrary parameter. Then there can be at most a probability ϵ^2 that the state is k and the sender has less than $\bar{\mu} - \Delta$ data distributed like f_k. A sender under state j has more than $\frac{\bar{\mu}-\Delta}{B}$ data with probability no more than $\frac{\Delta^2\epsilon^2 B^2}{(\bar{\mu}(1-B)-\Delta)^2}$.

Recall that whenever $u_{\sigma^*}(\mu,\theta) < \theta$, the type with dataset μf_θ is truthful in equilibrium. So, if under state θ_k we have $Pr[u_{\sigma^*}(\mu,\theta) < \theta - \delta] \geq \epsilon$, then the type with $\mu = G^{-1}(\epsilon)$ must obtain payoff less than $\theta - \delta$, and so must all types with less data, and all such types must be truthful. But the total mass of all types *not* in state k that can pool with types with $\mu \in [G^{-1}(\epsilon^2), G^{-1}(\epsilon)]$ cannot exceed

$$(J-1)(1-\beta_0(\theta_k))\frac{\Delta^2\epsilon^2B^2}{(\bar{\mu}(1-B)-\Delta)^2}$$

and so the payoff to type $G^{-1}(\epsilon)f_k$ cannot be less than

$$\frac{\epsilon(1-\epsilon)\theta_k}{\epsilon(1-\epsilon)+(J-1)(1-\beta_0(\theta_k))\frac{\Delta^2\epsilon^2B^2}{(\bar{\mu}(1-B)-\Delta)^2}}$$

which, for small enough Δ, must be at least $\theta_k - \delta$. Since there is always L large enough that $Var_L < \epsilon^2\Delta^2$, we are done.

All that remains is to note that, since the ex-ante expected payoff must always be $\mathbb{E}_{\beta_0}[\theta]$, this lower bound on the probability of payoffs less than the full-information payoffs implies a corresponding upper bound on payoffs exceeding the full-information payoffs, and so we obtain convergence of the distribution of payoffs, state-by-state, to those in the outcome where the receiver knows the truth. □

Comparative statics of welfare with respect to $\beta_0(\theta_j)$

Proof of Claim 5. First, let $M(v)$ be the frontier of types that attain payoff v under $\mathcal{G}$ and let $M'(v)$ be the frontier of types that do so under $\mathcal{G}'$. Let q be the distribution of types in $\mathcal{G}$ and q' be the type distribution for $\mathcal{G}'$.

Let $v \geq \theta_j$. Suppose for the sake of contradiction that $U(M') \setminus U(M)$ is nonempty. By Lemma 15, in the game $\mathcal{G}'$,

$$\mathbb{E}_{q'}[\theta | t \in U(M') \setminus U(M)] \geq v.$$

But we also have $\mathbb{E}_q[\theta | t \in U(M') \setminus U(M)] \geq \mathbb{E}_{q'}[\theta | t \in U(M') \setminus U(M)]$. So then $\mathbb{E}_{q'}[\theta | t \in U(M') \setminus U(M)] \geq v$, but then the construction algorithm in game $\mathcal{G}$, if it ever reached M, would instead set M' as a frontier for payoff v, and so this is impossible.

Similarly, let $v \leq \theta_j$. As with the above, we observe that if $U(M) \setminus U(M')$ is nonempty, then

$$\mathbb{E}_q[\theta | t \in U(M) \setminus U(M')] \geq v,$$

but since $\mathbb{E}_{q'}[\theta | t \in U(M) \setminus U(M')] \geq \mathbb{E}_q[\theta | t \in U(M) \setminus U(M')]$, this implies that $\mathbb{E}_q[\theta | t \in U(M) \setminus U(M')] \geq v$, which is likewise impossible by the algorithm. $\square$

A.3 Results on experimental design

Proposition 18. *Suppose that two games $\mathcal{G}$ and $\mathcal{G}'$ are identical except for their space of outcomes $\mathcal{D}$ and $\mathcal{D}'$ and the generating distributions of data under each state, $\{f_j\}_{j=1}^J$ and $\{f'_j\}_{j=1}^J$, and let σ^* and $\sigma^{*'}$ be their respective imitation equilibria.*

If the $r_j(k) \geq r'_j(k)$ for all j, k, then the receiver's payoff is greater under σ^* than under $\sigma^{*'}$.

Proof. Under game $\mathcal{G}$, there exists a (pure-strategy) mechanism that implements the outcome of $\sigma^{*'}$. To see this, note that the outcome of $\sigma^{*'}$ is also the outcome of v', the optimal mechanism for the receiver in $\mathcal{G}'$, which respects the IC constraints that can be rewritten as

$$v'(\mu f_j) \geq v'(\frac{\mu}{r_j(k)} f_k) \ \forall \mu, j, k \ \text{ and } v'(\mu_1 f_j) \geq v'(\mu_2 f_j) \ \forall j, \mu_1 > \mu_2. \tag{IC-$\mathcal{G}'$}$$

On the other hand, in order to be implementable in $\mathcal{G}$, v' need only respect the IC constraints

$$v'(\mu f_j) \geq v'(\frac{\mu}{r_j(k)} f_k) \ \forall \mu, j, k \ \text{ and } v'(\mu_1 f_j) \geq v'(\mu_2 f_j) \ \forall j, \mu_1 > \mu_2, \tag{IC-$\mathcal{G}$}$$

which are weaker.

Since v' is implementable in $\mathcal{G}$, the outcome of the optimal mechanism, and therefore the imitation equilibrium, in $\mathcal{G}$ gives at least a weak improvement over v' for the receiver. □

A.4 Properties of truth-leaning equilibria with finite data

Claim 19. *In any finite-data game $\mathcal{G}_N$, the unique inclusive announcement-proof equilibrium outcome is the truth-leaning equilibrium outcome.*

To show that the truth-leaning equilibrium outcome is inclusive announcement-proof in finite-data games, I construct it, using the algorithm from Rappoport, which I summarize here. In short, the equilibrium is constructed by iteratively choosing a frontier of types such that the set of types "above" the frontier, in the sense of being able to imitate some frontier type, yields as favorable a belief as possible.

Algorithm (Finite N). First, define for any type set T the subset of types $T^+(M) = T \bigcap U(M)$ as the set of types in T that are capable of sending some message in message set M, and define $u_{pool}(T)$ to be the payoff to the sender if the receiver knows only that their type must be in T.

1. Let $T_1 = \mathcal{T}_N$, and find the set of messages $M_1 \subseteq T_1$ that maximizes the payoff to a pool consisting of the set of senders in T_1 who can send at least one message in it:

$$M_1 \in \arg\max_{M \subseteq T_1} u_{pool}(T_1^+(M)).$$

   If there are multiple such pools, then we take their union, which is also such a pool.

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)

$$M_s \in \arg\max_{M \in T_s} u_{pool}(T_s^+(M)).$$

3. Continue until $T_s \setminus T_s^+(M_s) = \emptyset$. Given each set M_s, there always exists a mixed strategy profile σ_{pool}^M defined over types in $T_1^+(M)$ such that each message in M yields the same payoff under the receiver's induced beliefs from σ_{pool}^M.[2] Define σ^* by $\sigma^*(m|t) = \hat{\sigma}_{pool}^{M_s}(m|t)$ where M_s is the pool containing m.

Proof. (Unique credible inclusive announcement-proof outcome). By construction, there is no credible inclusive announcement, since such an announcement would constitute a better set of types than the one constructed at some step of the algorithm; this violates the optimality of the pool of types constructed in each step. No other outcome is immune: 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^*. □

[2] Otherwise, the worst possible payoff to particular message in M over all strategy profiles over M_s is better than the best possible payoff to some other message; then there always exists $M \subset M_s$ such that $T_s^+(M) > T_s^+(M_s)$.

A.5 Proof of convergence to imitation equilibrium as $N \to \infty$

Here, we first use the separation theorem to prove that outcomes $u_{\sigma^*_N}$ converge to u_{σ^*} in a convergent sequence of games; then, we state a corresponding result that describes the extent to which the sender's messaging strategy also converges, and we give a proof.

To prove Theorem 6, we first introduce some notation. Let payoff frontiers in $\mathcal{G}_N$ under a truth-leaning equilibrium σ^*_N be given by $\hat{\mathbf{M}}_N(v)$ — that is,

$$\hat{\mathbf{M}}_N(v) = \{t \in \mathcal{T}_N : u_{\sigma^*_N}(t) \geq v, \ \nexists t' \in \mathcal{T}_N \text{ w/ } t' \hat{\subset} t \text{ and } u_{\sigma^*_N}(t') \geq v\}.$$

Let a translation of $M_N(v)$ to the restricted limit type space $\mathcal{T}$ be $\hat{\mu}^N(v)$, where

$$\hat{\mu}^N_j(v) = \{\mu \in [0,1] : u_{\sigma^*_N}(\mu f_j) \geq v, \text{ and } u_{\sigma^*_N}(\mu' f_j) < v \ \forall \mu' < \mu\}.$$

Finally, let

$$u_{pool,N}(T) = \frac{\sum_{t \in T} q_N(t) \mathbb{E}_{\pi(\cdot|t)}[\theta]}{\sum_{t \in T} q_N(t)}$$

be the analog of u_{pool} for finite game $\mathcal{G}_N$. We first give a lemma that shows that average values in the finite game converge to those in the continuous-μ game within upper and lower pools.

Lemma 19. *Fix some $\epsilon > 0$ and $\delta > 0$. We aim to show that there exists large-enough $\hat{N}(\epsilon, \delta)$ such that for $N > \hat{N}(\epsilon, \delta)$, we can ensure that neither $U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$ nor $U_N(\hat{\mathbf{M}}_N(v)) \setminus U(\hat{\mu}(v - \epsilon))$ contain more than a measure δ of types in $\mathcal{T}$.*

Proof of Lemma 19. Fix an integer n. The Glivenko-Cantelli theorem implies that there is a bound on the probability that $\sup_d |\sum_{x=1}^d t(x) - \frac{n}{N} F_j(d)| > \eta$ conditional on $|t| = n$ and the true state being θ_j 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. Before proceeding further, we formalize the implications for the problem at hand.

It is helpful to formally define a neighborhood of $\{\mu f_j : \mu \in [0,1]\}$ in $\mathcal{T}_N$. For $\eta > 0$, define

$$S^j_N(\eta) = \{t \in \mathcal{T}_N : \exists \theta \text{ s.t. } \sup_d |t(d) - |t| f_\theta(d)| \leq \eta\}.$$

This is the set of datasets in $\mathcal{T}_N$ that differ from some type in $\mathcal{T}$ for which the state is θ_j by no more than a fraction η of observations of each outcome. Furthermore, taking any lower bound $k^* \in (0,1]$, let $S^j_N(\eta, k^*)$ be the set $S^j_N(\eta) \bigcap \{t : |t| \geq N k^*\}$ of all datasets in the neighborhood that contain at least a fraction k^* of total possible observations.

We can ensure that, if we know that the state is $\theta = theta_j$ and the number of observations of the data the sender observes is $n > N k^*$, then the likelihood that their type lies in $S^j_N(\eta, k^*)$ is close to 1 for all η as long as N is sufficiently large. Furthermore, if η is small enough for a

given value of k^*, then $S_N^j(\eta, k^*)$ are disjoint. There is a sufficient upper-bound to the value of η that achieves this, $\tilde{\eta}(k^*) = \min_{j,j'} \max_d k\frac{f_j(d)-f_{j'}(d)}{2}$.

For any desired likelihoods $l_1 > 0$ and $l_2 > 0$, and any error rates $\xi_1 > 0$ and $\xi_2 > 0$, there exists a uniform bound $\tilde{N}(k^*, \eta, l_1, l_2, \xi_1, \xi_2)$ such that as long as $k > k^*$ and $N > \tilde{N}(k^*, \eta, l_1, l_2, \xi_1, \xi_2)$, with $\eta \leq \tilde{\eta}(k^*)$,

a) $Pr(t \in S_N^j(\eta, k^*)|\theta = \theta_j, n = Nk) \geq 1 - l_1$

b) $Pr(t \in S_N^j(\eta, k^*)|\theta \neq \theta_j, n = Nk) < l_2$ for any $k > k^*$

c) $|\mathbb{E}_{\pi_N(\cdot|t)}[\theta] - \theta_j| \leq \xi_1$ for all $t \in S_N^j(\eta, k^*)$

d) $|(G_N(Nk) - G_N(Nk')) - (G(k) - G(k'))| < \xi_2|$ for all $0 \leq k' \leq k \leq 1$.

Part (a) follows directly from the Glivenko-Cantelli theorem. Part (b) follows from applying part (a) to $j' \neq j$, although the bound could certainly be tightened more. Part (c) follows from both the previous parts and the fact that the set of possible values of θ is finite. Part (d) follows from the convergence of $\mathcal{G}_1, \ldots,$ to $\mathcal{G}$.

Now we can bound the average value of $U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{M}(v))$ in $\mathcal{T}_N$.

Let us start with types in $S_N^j(\eta, k) \bigcap U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{M}(v))$. In $\mathcal{T}$, we know that $U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{M}(v))$ contains μf_j for $\mu \in [\hat{\mu}_j(v + \epsilon), \hat{\mu}_j^N(v)]$. Recall that $|\mathcal{D}| = D$, and define $R = \max_{j,d,d'} \frac{f_j(d)}{f_j(d')}$. We know that for all $k \in [\hat{\mu}_j(v + \epsilon) + DR\eta, \hat{\mu}_j^N(v) - DR\eta]$, if $t \in S_N^j(\eta, k)$ and $|t| = Nk$ then $t \in S_N^j(\eta, k) \bigcap U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$. This is true because:

- If $t \in S_N^j(\eta, k)$, then the nearest type $\mu f_j \in \mathcal{T}$ differs by adding or deleting at most a mass η of observations of each outcome.
- As a result, the type $(k + DR\eta)f_j$ can imitate t, and the type $(k - DR\eta)f_j$ can be imitated by t.
- Therefore, if, in addition, $k \in [\hat{\mu}_j(v + \epsilon) + RD\eta, \hat{\mu}_j^N(v) - RD\eta]$, then $t \in U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$.

Likewise, unless $k \in [\hat{\mu}_j(v + \epsilon) - DR\eta, \hat{\mu}_j^N(v) + DR\eta]$, if $t \in S_N^j(\eta, k)$ and $|t| = Nk$ then $t \notin S_N^j(\eta, k) \bigcap U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$.

The probability that the raw dataset in $\mathcal{G}_N$ lies in $U(\hat{\mu}(v + \epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$ is therefore lower-bounded by

$$\beta_0(\theta_j)(1 - l_1)[G(\hat{\mu}_j^N(v) + RD\eta(k^*)) - G(\hat{\mu}_j(v + \epsilon) - RD\eta(k^*)) - 2\xi_2] - G_N(k^*N)$$

and upper-bounded by

$$\beta_0(\theta_j)[G(\hat{\mu}_j^N(v) - RD\eta(k^*)) - G(\hat{\mu}_j(v + \epsilon) + RD\eta(k^*)) + 2\xi_2] + l_2$$

whenever $N > \tilde{N}(k^*, \eta, l_1, l_2, \xi_1, \xi_2)$.

In addition, we can define the upper bound $\bar{b} = \max_\mu g(\mu)$ because g is continuous on a compact interval. Then if

$$\sum_{j=1}^{J}[G(\hat{\mu}_j^N(v)) - G(\hat{\mu}_j(v+\epsilon))] \geq \delta,$$

a crude lower bound on the average value of types in $U(\hat{\mu}(v+\epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$ under the finite-game type distribution q_N is

$$\begin{aligned}
&\frac{\sum_{j=1}^{J}[\beta_0(\theta_j)(1-l_1)[G(\hat{\mu}_j^N(v)+RD\eta) - G(\hat{\mu}_j(v+\epsilon)-RD\eta) - 2\xi_2] - G(k) - \xi_2](\theta_j - \xi)}{\beta_0(\theta_j)[G(\hat{\mu}_j^N(v)-RD\eta) - G(\hat{\mu}_j(v+\epsilon)+RD\eta) + 2\xi_2] + l_2} \\
&\leq \frac{\delta(v-\xi) - l_1 - 2\bar{b}RD\eta - G(k) - 3\xi_2}{\delta + l_2 + l_1 + 2\bar{b}RD\eta + 2\xi_2} \\
&\equiv LB(k^*, \eta, l_1, l_2, \xi_1, \xi_2|\delta)
\end{aligned} \tag{A.6}$$

as long as $N > \tilde{N}(k^*, \eta, l_1, l_2, \xi_1, \xi_2)$. Note that this bound is independent of v, that is, it applies uniformly to all payoff frontiers.

We have that

$$\lim_{N\to\infty} \min_{\substack{k^*,\eta,l_1,l_2,\xi_1,\xi_2: \\ \tilde{N}(k^*,\eta,l_1,l_2,\xi_1,\xi_2)\leq N}} LB(\eta, k^*, l_1, l_2, \xi_1, \xi_2|\delta) = \frac{\sum_{j=1}^{J}[G(\hat{\mu}_j^N(v)) - G(\hat{\mu}_j(v+\epsilon))]\theta_j}{\sum_{j=1}^{J}[G(\hat{\mu}_j^N(v)) - G(\hat{\mu}_j(v+\epsilon))]} \geq v + \epsilon,$$

by the separation theorem applied to types in $U(\hat{\mu}(v+\epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$ under the continuous-data type distribution q. But we also know by applying the separation theorem that the average value of types in $U(\hat{\mu}(v+\epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$ under the finite-game type distribution q_N is no more than v. Then there exists large-enough $\hat{N}_+(\epsilon, \delta)$ such that for $N > \hat{N}_+(\epsilon, \delta)$, the above bound ensures that for any v, the set $U(\hat{\mu}(v+\epsilon)) \setminus U_N(\hat{\mathbf{M}}_N(v))$ does not contain more than a measure δ of types in $\mathcal{T}$.

We can show, using a completely symmetric argument, that there also exists large-enough $\hat{N}_-(\epsilon, \delta)$ such that for $N > \hat{N}_-(\epsilon, \delta)$, the set $U_N(\hat{\mathbf{M}}_N(v)) \setminus U(\hat{\mu}(v-\epsilon))$ does not contain more than a measure δ of types in $\mathcal{T}$, regardless of v. Rather than lower-bounding the average value of all types in $U_N(\hat{\mathbf{M}}_N(v)) \setminus U(\hat{\mu}(v-\epsilon))$ under q_N, we upper-bound it and show that given fixed δ, for large N it is less than ϵ greater than the value of the same set of types evaluated under type distribution q. This shows that it has average value less than v, and fails the separation theorem for the finite game. □

Next, we use the the lemma to prove the theorem.

Proof of Theorem 6. There is an upper bound $\bar{\Delta} = \max_{\mu,j} \frac{d\hat{u}_j(\mu)}{d\mu}$ on the rate of change of payoffs in μ under σ^*.

In addition, for every $y > 0$, there is some $C(y) = \min_{\mu,j} \beta_0(\theta_j)(G(\mu) - G(\mu - y)) > 0$ with $\lim_{x\to 0} C^{-1}(x) = 0$, which shows that upper-bounding the measure of an interval $[\mu - y, \mu]$ of measures of types under state θ_j with respect to prior distribution q also upper-bounds y itself.

Then we know that for a type $\mu f_j \in T$ with $u_{\sigma^*}((\mu + C^{-1}\delta)f_j) = v$, we have

$$u_{\sigma^*_N}(\mu f_j) \leq u_{\sigma^*}((\mu + C^{-1}\delta)f_j) + \epsilon \leq u_{\sigma^*}(\mu f_j) + C^{-1}(\delta)\bar{\Delta} + \epsilon$$

and likewise

$$u_{\sigma^*_N}(\mu f_j) \geq u_{\sigma^*}((\mu - C^{-1}\delta)f_j) - \epsilon \geq u_{\sigma^*}(\mu f_j) - C^{-1}(\delta)\bar{\Delta} - \epsilon$$

for all $N > \hat{N}(\epsilon, \delta)$.

We can take

$$\min_{\delta,\epsilon: N > \hat{N}(\epsilon,\delta)} C^{-1}(\delta)\bar{\Delta} + \epsilon$$

to be the bound on the difference between $u_{\sigma^*}(t) - u^*_{\sigma^*_N}(t)$, and it shrinks to 0 as $N \to \infty$. □

A.5.1 Strategic convergence

Proposition 20. *Suppose that $\{\mathcal{G}_N\}_{N=1}^{\infty}$ converge to $\mathcal{G}_\infty$. Then for all $p^*, \rho, \eta > 0$, there is $\underline{N}(p^*, \eta)$ such that for all $N > \underline{N}(p^*, \eta)$, conditional on $|u_{\sigma^*_N}(t) - \theta_k| > \eta$ for all k, there is at least probability $1 - p^*$ that t sends a message with (sup norm) distance at most δ from some $t_\infty \in \mathcal{T}_\infty$ that is on-path in σ^*_∞ and such that $|u_{\sigma^*_N}(t) - u_{\sigma^*_\infty}(t_\infty)| \leq \rho$.*

This is a partial characterization of large-N equilibrium strategies, saying that among types that obtain payoff bounded away by an arbitrarily small amount from the rewards to certainty about any particular state, the likelihood of playing a message close to their optimal message under limit-game beliefs is very high when there is plentiful access to data. In other words, these types play imitation-like strategies. This follows from the convergence theorem, since in truth-leaning equilibrium a type that receives payoff less than its full-information payoff always discloses its full dataset, so types with $u_{\sigma^*_N}(t) << \mathbb{E}_{\pi(\cdot|t)}[\theta]$ tell the truth; convergence of outcomes implies that they receive payoffs similar to those obtained by nearby types in T_∞ under $\beta_{\sigma^*_\infty}$, and convergence of the type distribution implies that most such senders are indeed near some type in T_∞. In aggregate a similar set of imitators must pool with such senders as the set of imitators pooling with better-state senders in σ^*_∞, which means that types with $u_{\sigma^*_N}(t) >> \mathbb{E}_{\pi(\cdot|t)}[\theta]$ play messages close to $\mathcal{T}_\infty$ with high probability.

The caveat is that when $|\mathbb{E}_{\beta_{\sigma^*_N}(\cdot|m)}[\theta] - \theta_k|$ is small for some k, then there may be no significant mass of senders playing m to earn a payoff much greater or much less than their full-information payoff, which makes it hard to apply the technique of matching imitators

to the imitated, though we do not have a counterexample for this case. From Corollary 1.4, we know that there is a positive-measure set of types that receive payoffs close to their full-information payoffs in σ^*_∞, and the proposition does not pin down the large-N limit of equilibrium strategies of types close to them, but generically, besides these, the set of types excluded from the proposition is measure-0.[3]

Proof of Prop. 20. Define $m_{\sigma^*_N}(t)$ to be the realization of the message played when the sender's type is t – formally, $m_{\sigma^*_N}(t)$ is a random variable with outcomes in $\mathcal{M}_N$ whose distribution is given by the equilibrium strategy $\sigma^*_N(\cdot|t)$.

Define $A_N(x, \Delta; \epsilon)$ to be the set of types $t \in \mathcal{T}_N \bigcap T(\epsilon)$ such that $u_{\sigma^*_N}(t) > \mathbb{E}_{\pi(\cdot|t)}[\theta]$, and $u_{\sigma^*_N}(t) \in (x, x + \Delta]$.

Define $B_N(x, \Delta; \epsilon)$ to be the set of types $t' \in \mathcal{T}_N \bigcap T(\epsilon)$ with $u_{\sigma^*_N}(t') < \mathbb{E}_{\pi(\cdot|t)}[\theta]$, and $u_{\sigma^*_N}(t') \in (x, x + \Delta]$.

Define $X(\eta, \xi, \omega) = \{x \in \left[\max_j \hat{u}_j(\xi) + \omega, u_{\sigma^*}(f_{\theta_J})\right) : \min_k |u_{\sigma^*_N}(t) - \theta_k| > \eta\}$.

For small enough η, ξ and ω, $X(\eta, \xi, \omega)$ is nonempty. On the other hand, $A_N(x, \Delta; \epsilon)$ and $B_N(x, \Delta; \epsilon)$ may be empty, in particular for small N. However, for $x \in X(\eta, \xi, \omega)$, there is large-enough $\underline{N}^*$ so that they are nonempty for all $N > \underline{N}^*$. Continuity of $u_{\sigma^*}(\mu f_j)$ ensures there is positive-measure set of types in $\mathcal{T}$ with $u_{\sigma^*_{infty}}(t) \in [x, x + \Delta]$; the bound away from θ_k for all k ensures that some such types have $u_{\sigma^*_{infty}}(t) - \mathbb{E}_{\pi(\cdot|t)}[\theta] \geq \eta$ and some have $u_{\sigma^*_i nfty}(t) \leq \mathbb{E}_{\pi(\cdot|t)}[\theta] < -\eta$, and so there is a positive-measure set of types nearby with the same properties under σ^*_N in $\mathcal{T}_N$ for large-enough N.

We first prove a claim.

Claim 20. *If $\{\sigma^*_N\}_{N=1}^\infty$ are truth-leaning equilibria of games $\mathcal{G}_N$ that converge to limit game $\mathcal{G}$ with imitation equilibrium σ^*, then for any $\eta > 0$, $\xi > 0$, $\omega > 0$ and $p > 0$, there exists $\bar{\epsilon} > 0$ and $\bar{\Delta} > 0$ such that, for all $x \in X(\eta, \xi, \omega)$, the probability conditional on $t \in A_N(x, \Delta; \epsilon)$ that $m_{\sigma^*_N}(t) \in B_N(x, \Delta; \epsilon)$ is at least $1 - p$ in the limit as $N \to 0$ for all $\epsilon < \bar{\epsilon}$ and $\Delta < \bar{\Delta}$.*

Proof of Claim 20. Expanding out the realization of $m_{\sigma^*_N}(t)$, this is equivalent to saying that for given $\eta > 0$, $\xi > 0$, $\omega > 0$, $p > 0$, there exists $\bar{\Delta} > 0$ and $\bar{\epsilon} > 0$ so that for all $\Delta < \bar{\Delta}$ and $\epsilon < \bar{\epsilon}$,

$$\lim_{N \to \infty} \frac{\sum_{t \in A_N(x,\Delta;\epsilon)} \left[q_N(t) \sum_{t' \in B_N(x,\Delta;\epsilon)} \sigma^*_N(t'|t)\right]}{\sum_{t \in A_N(x,\Delta;\epsilon)} q_N(t)} \geq 1 - p.$$

for all $x \in X(\eta, \xi, \omega)$.

[3] Genericity here can be with respect to perturbations in β_0 or $\theta_1, \ldots, \theta_J$.

We have that for any $\xi > 0$, $\omega > 0$,

$$\lim_{\epsilon \to 0} \lim_{N \to \infty} \min_{\theta_j} \max_{t \in \mathcal{T}_N \bigcap T(\epsilon): u_{\sigma_N^*}(t) \geq \max_j \hat{u}_j(\xi) + \omega} |\theta_j - \mathbb{E}_{\pi(\cdot|t)}[\theta]| = 0.$$

Thus for any $\nu > 0$, $\omega > 0$ and $\xi > 0$, there exist small-enough $\bar{\epsilon}(\nu, \xi, \omega) > 0$ and large-enough $\underline{N}(\nu, \xi, \omega, \epsilon)$ defined for $\epsilon < \bar{\epsilon}(\nu, \xi)$ such that $\min_{\theta_j} \max_{t \in \mathcal{T}_N \bigcap T(\epsilon): u_{\sigma_N^*}(t) \geq \max_j \hat{u}_j(\xi) + \omega} |\theta_j - \mathbb{E}_{\pi(\cdot|t)}| < \nu$ for all $\epsilon < \bar{\epsilon}(\nu, \xi, \omega)$, $N > \underline{N}(\nu, \xi, \omega, \epsilon)$.

If we take $\Delta < \eta/3$ and $\nu < \eta/3$, then whenever $|\theta_k - \mathbb{E}_{\pi(\cdot|t)}| < \nu$ for some k and $u_{\sigma_N^*}(t) \in [x, x + \Delta]$ for x in $X(\eta, \xi, \omega)$, we have that $|u_{\sigma_N^*}(t) - \mathbb{E}_{\pi(\cdot|t)}| > \eta/3$. In particular, $u_{\sigma_N^*}(t) \neq \mathbb{E}_{\pi(\cdot|t)}$, so, for any $x \in X(\eta, \xi, \omega)$, and for any $\Delta, \nu < \eta/3$ and any ξ, ω and $\epsilon < \bar{\epsilon}(\nu, \xi, \omega)$, $N > \underline{N}(\nu, \xi, \omega, \epsilon)$,

$$A_N(x, \Delta; \epsilon) \bigcup B_N(x, \Delta; \epsilon) = \{t \in T(\epsilon) \bigcap \mathcal{T}_N : u_{\sigma_N^*}(t) \in (x, x + \Delta]\}.$$

In addition, the uniform convergence of outcomes on $\mathcal{T}$ and of the type distribution conditional on each state ensures that for all ϵ, ξ, Δ, and for all $x \geq \max_j \hat{u}_j(\xi) + \omega$,

$$\lim_{N \to \infty} \sum_{t \in A_N(x, \Delta; \epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta] = \sum_{\theta_j < \theta_k} \theta_j \beta_0(\theta_j) [G^j(\hat{\mu}_j(x + \Delta)) - G^j(\hat{\mu}_j(x))] \tag{A.7}$$

and

$$\lim_{N \to \infty} \sum_{t \in A_N(x, \Delta; \epsilon)} q(t) = \sum_{\theta_j < \theta_k} \beta_0(\theta_j) [G^j(\hat{\mu}_j(x + \Delta)) - G^j(\hat{\mu}_j(x))] \tag{A.8}$$

and likewise

$$\lim_{N \to \infty} \sum_{t \in B_N(x, \Delta; \epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta] = \sum_{\theta_j > \theta_k} \theta_j \beta_0(\theta_j) [G^j(\hat{\mu}_j(x + \Delta)) - G^j(\hat{\mu}_j(x))] \tag{A.9}$$

and

$$\lim_{N \to \infty} \sum_{t \in B_N(x, \Delta; \epsilon)} q(t) = \sum_{\theta_j > \theta_k} \beta_0(\theta_j) [G^j(\hat{\mu}_j(x + \Delta)) - G^j(\hat{\mu}_j(x))]. \tag{A.10}$$

Supposing that $x \in X(\eta, \xi, \omega)$ for some θ_k, and $\epsilon < \bar{\epsilon}(\nu, \xi, \omega)$, and $\Delta, \nu < \eta/3$, we know from the above that

$$\begin{aligned} x &\leq \frac{\sum_{\theta_j} \theta_j \beta_0(\theta_j) [G^j(\hat{\mu}_j(x + \Delta)) - G^j(\hat{\mu}_j(x))]}{\sum_{\theta_j} \beta_0(\theta_j) [G^j(\hat{\mu}_j(x + \Delta)) - G^j(\hat{\mu}_j(x))]} \\ &= \lim_{N \to \infty} \frac{\sum_{t \in B_N(x, \Delta; \epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta] + \sum_{t \in A_N(x, \Delta; \epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta]}{\sum_{t \in B_N(x, \Delta; \epsilon)} q(t) + \sum_{t \in A_N(x, \Delta; \epsilon)} q(t)} \end{aligned} \tag{A.11}$$

On the other hand,

$$
\begin{aligned}
&x+\Delta \\
&\geq \lim_{N\to\infty} \frac{\sum_{t\in B_N(x,\Delta;\epsilon)} q(t)\mathbb{E}_{\pi(\cdot|t)}[\theta] + \sum_{t\in A_N(x,\Delta;\epsilon)} \left[q(t)\mathbb{E}_{\pi(\cdot|t)}[\theta] \sum_{t'\in B_N(x,\Delta;\epsilon)} \sigma_N^*(t'|t)\right]}{\sum_{t\in B_N(x,\Delta;\epsilon)} q(t) + \sum_{t\in A_N(x,\Delta;\epsilon)} \left[q(t) \sum_{t'\in B_N(x,\Delta;\epsilon)} \sigma_N^*(t'|t)\right]} \\
&= \lim_{N\to\infty} \frac{\sum_{\theta_j} \theta_j\beta_0(\theta_j)[G^j(\hat{\mu}_j(x+\Delta)) - G^j(\hat{\mu}_j(x))] - \sum_{t\in A_N(x,\Delta;\epsilon)} q(t)\mathbb{E}_{\pi(\cdot|t)}[\theta] \sum_{t'\in B_N(x,\Delta;\epsilon)}(1-\sigma_N^*(t|t'))}{\sum_{\theta_j} \beta_0(\theta_j)[G^j(\hat{\mu}_j(x+\Delta)) - G^j(\hat{\mu}_j(x))] - \sum_{t\in A_N(x,\Delta;\epsilon)} q(t) \sum_{t'\in B_N(x,\Delta;\epsilon)}(1-\sigma_N^*(t|t'))}.
\end{aligned} \tag{A.12}
$$

But, we know that $\mathbb{E}_{\pi(\cdot|t)}[\theta] < u_{\sigma_N^*}(t) - \eta/3 \leq x+\Delta-\eta/3$ for any $t \in A_N(x,\Delta;\epsilon)$. Then, combining, we have

$$
\begin{aligned}
0 \leq& (x+\Delta)\left(\sum_{\theta_j} \beta_0(\theta_j)[G^j(\hat{\mu}_j(x+\Delta)) - G^j(\hat{\mu}_j(x))] - \lim_{N\to\infty} \sum_{t\in A_N(x,\Delta;\epsilon)} q(t) \sum_{t'\in B_N(x,\Delta;\epsilon)} (1-\sigma_N^*(t|t'))\right) \\
&- x\left(\sum_{\theta_j} \beta_0(\theta_j)[G^j(\hat{\mu}_j(x+\Delta)) - G^j(\hat{\mu}_j(x))]\right) - \lim_{N\to\infty} \sum_{t\in A_N(x,\Delta;\epsilon)} q(t)\mathbb{E}_{\pi(\cdot|t)}[\theta] \sum_{t'\in B_N(x,\Delta;\epsilon)} (1-\sigma_N^*(t|t')) \\
\leq& \Delta\left(\sum_{\theta_j} \beta_0(\theta_j)[G^j(\hat{\mu}_j(x+\Delta)) - G^j(\hat{\mu}_j(x))]\right) \\
&\qquad - [(x+\Delta) - (x+\Delta-\eta/3)] \lim_{N\to\infty} \sum_{t\in A_N(x,\Delta;\epsilon)} q(t) \sum_{t'\in B_N(x,\Delta;\epsilon)} (1-\sigma_N^*(t|t')) \\
=& \lim_{N\to\infty} \Delta\left(\sum_{t\in A_N(x,\Delta;\epsilon)\bigcup B_N(x,\Delta;\epsilon)} q(t)\right) - \frac{\eta}{3} \lim_{N\to\infty} \sum_{t\in A_N(x,\Delta;\epsilon)} q(t) \sum_{t'\in B_N(x,\Delta;\epsilon)} (1-\sigma_N^*(t|t')).
\end{aligned} \tag{A.13}
$$

Finally, since

$$
\lim_{N\to\infty} \frac{\sum_{t\in A_N(x,\Delta;\epsilon)\bigcup B_N(x,\Delta;\epsilon)} \mathbb{E}_{\pi(\cdot|t)}[\theta] q(t)}{\sum_{t\in A_N(x,\Delta;\epsilon)\bigcup B_N(x,\Delta;\epsilon)} q(t)} \leq x+\Delta,
$$

we have

$$
\sum_{t\in B_N(x,\Delta;\epsilon)} (\theta_{k+1} - x - \Delta) q(t) \leq \sum_{t\in A_N(x,\Delta;\epsilon)} (x+\Delta-\theta_1) q(t)
$$

and so

$$
\frac{\sum_{t\in A_N(x,\Delta;\epsilon)} q(t)}{\sum_{t\in B_N(x,\Delta;\epsilon)} q(t)} \geq \frac{\eta/3}{\theta_J}.
$$

From the above and eq. A.13, we have

$$\lim_{N\to\infty} \frac{\sum_{t\in A_N(x,\Delta;\epsilon)} \left[q_N(t) \sum_{t'\in B_N(x,\Delta;\epsilon)} (1-\sigma_N^*(t'|t))\right]}{\sum_{t\in A_N(x,\Delta;\epsilon)} q_N(t)} \leq \frac{\Delta(\theta_J+\eta)}{(\eta/3)^2}.$$

Then for given p, ξ, ω, η, and $\nu < \eta/3$ and $\epsilon < \bar{\epsilon}(\nu,\xi,\omega)$, as long as $\Delta \leq \frac{p\eta^2}{9(\theta_J+\eta)}$, we have

$$\lim_{N\to\infty} \frac{\sum_{t\in A_N(x,\Delta;\epsilon,\xi)} \left[q_N(t) \sum_{t'\in B_N(x,\Delta;\epsilon,\xi)} \sigma_N^*(t'|t)\right]}{\sum_{t\in A_N(x,\Delta;\epsilon,\xi)} q_N(t)} \geq 1-p$$

and this bound is independent of x. So, letting $\bar{\epsilon} = \bar{\epsilon}(\eta/3,\xi,\omega)$ and $\bar{\Delta} = \frac{p\eta^2}{9(\theta_J+\eta)}$, we have proven the claim. □

Next, let $\underline{u} = \hat{u}_j(0)$, where the choice of j for the definition does not matter. Suppose the following condition holds for some positive η:

Condition 1. $\min_k |\theta_k - \underline{u}| > \eta$ *and there is* $\tilde{\xi} > 0$ *such that* $G^j(\tilde{\xi}) > 0$ *for all* j, *and* $\hat{u}_j(\tilde{\xi}) = \underline{u}$ *for some* j.

This says that a sender with a positive amount $\tilde{\xi}$ of distribution f_j gets the same payoff as the sender with no data, and that that payoff is bounded away from any θ_j by η. When showing that senders must play similarly under σ_N^* in the limit as the average dataset becomes large, we consider separately the small fraction of senders that, by chance, receive very little data, i.e. those with $|t| \leq \xi$, and in this case

$$\lim_{\xi\to 0}\lim_{\omega\to 0} \min\{\mu : \exists j \text{ s.t. } \hat{u}_j(\mu) > \xi+\omega\} > 0,$$

and there is a positive-measure set of types that may be pooled with those low-data senders.

Let us prove a similar claim to the previous one.

Claim 21. *If there exists* η *such that condition 1 holds, then for given* $p > 0$, *there exists* $\bar{\epsilon} > 0$ *such that for* $\epsilon < \bar{\epsilon}$, *if we define*

$$S(\xi,\omega,\epsilon) = \{t \in \mathcal{T}_N \bigcap T(\epsilon) : |t| \leq \xi \text{ and } u_{\sigma_N^*}(t) \in (x, x+\Delta]\},$$

then letting $a_N(\xi,\omega,\epsilon) = A_N(\underline{u}-\omega, 2\omega;\epsilon)\backslash S(\xi,\omega,\epsilon)$ *and* $b_N(\xi,\omega,\epsilon) = B_N(\underline{u}-\omega,2\omega;\epsilon)\bigcup S(\xi,\omega,\epsilon)$, *we have*

$$\lim_{\xi\to 0}\lim_{\omega\to 0}\lim_{N\to\infty} \frac{\sum_{t\in a_N(\xi,\omega,\epsilon)} \left[q_N(t) \sum_{t'\in b_N(\xi,\omega,\epsilon)} \sigma_N^*(t'|t)\right]}{\sum_{t\in a_N(\xi,\omega,\epsilon)} q_N(t)} \geq 1-p.$$

Proof of Claim 21. To start, note that in this case, we have for all $\xi > 0$ that

$$\lim_{\epsilon \to 0} \lim_{N \to 0} \min_{\theta_j} \max_{t : |t| \leq \xi} |\theta_j - \mathbb{E}_{\pi(\cdot|t)}[\theta]| = 0.$$

Then for all $t \in a_N(\xi, \omega, \epsilon)$, there is some $\bar{\epsilon}'(\eta, \xi)$ and $\underline{N}'(\eta, \xi, \epsilon)$ so that for all $\epsilon < \bar{\epsilon}'(\eta, \xi)$ and $N > \underline{N}'(\eta, \xi, \epsilon)$, we have $u_{\sigma_N^*}(t) - \mathbb{E}_{\pi(\cdot|t)}[\theta] \geq \eta/3$.

We know that

$$\begin{aligned} \underline{u} - \omega &\leq \lim_{N \to \infty} \frac{\sum_{t \in b_N(\xi,\omega,\epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta] + \sum_{t \in a_N(\xi,\omega,\epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta]}{\sum_{t \in b_N(\xi,\omega,\epsilon)} q(t) + \sum_{t \in a_N(\xi,\omega,\epsilon)} q(t)} \\ &= \frac{\sum_{\theta_j} \theta_j \beta_0(\theta_j) G^j(\hat{\mu}_j(\underline{u} + \omega))}{\sum_{\theta_j} \beta_0(\theta_j) G^j(\hat{\mu}_j(\underline{u} + \omega))} \end{aligned} \tag{A.14}$$

and

$$\begin{aligned} \underline{u} + \omega &\geq \lim_{N \to \infty} \frac{\sum_{t \in b_N(\xi,\omega,\epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta] + \sum_{t \in a_N(\xi,\omega,\epsilon)} \left[q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta] \sum_{t' \in b_N(\xi,\omega,\epsilon)} \sigma_N^*(t'|t) \right]}{\sum_{t \in b_N(\xi,\omega,\epsilon)} q(t) + \sum_{t \in a_N(\xi,\omega,\epsilon)} \left[q(t) \sum_{t' \in b_N(\xi,\omega,\epsilon)} \sigma_N^*(t'|t) \right]} \\ &= \lim_{N \to \infty} \frac{\sum_{\theta_j} \theta_j \beta_0(\theta_j) G^j(\hat{\mu}_j(\underline{u} + \omega)) - \sum_{t \in a_N(\xi,\omega,\epsilon)} q(t) \mathbb{E}_{\pi(\cdot|t)}[\theta] \sum_{t' \in b_N(\xi,\omega,\epsilon)} (1 - \sigma_N^*(t|t'))}{\sum_{\theta_j} \beta_0(\theta_j) G^j(\hat{\mu}_j(\underline{u} + \omega)) - \sum_{t \in a_N(\xi,\omega,\epsilon)} q(t) \sum_{t' \in b_N(\xi,\omega,\epsilon)} (1 - \sigma_N^*(t|t'))}. \end{aligned} \tag{A.15}$$

Then, just as in eq. A.13, we have when $\epsilon < \bar{\epsilon}'(\eta, \xi, \omega)$ that

$$\lim_{N \to \infty} 2\omega \left(\sum_{t \in a_N(\xi,\omega,\epsilon) \bigcup b_N(\xi,\omega,\epsilon)} q(t) \right) - \frac{\eta}{3} \lim_{N \to \infty} \sum_{t \in a_N(x,\Delta;\epsilon)} q(t) \sum_{t' \in B_N(x,\Delta;\epsilon)} (1 - \sigma_N^*(t|t')) \geq 0.$$

Since there is some j such that $\hat{u}_j(\tilde{\xi}) = \underline{u}$, we have the bound

$$\lim_{N \to \infty} \sum_{t \in a_N(\xi,\omega,\epsilon)} q_N(t) \geq \beta_0(\theta_j)[G^j(\tilde{\xi}) - G^j(\xi)].$$

So, we have that

$$\lim_{N \to \infty} \frac{\sum_{t \in a_N(\xi,\omega,\epsilon)} \left[q_N(t) \sum_{t' \in b_N(\xi,\omega,\epsilon)} (1 - \sigma_N^*(t'|t)) \right]}{\sum_{t \in a_N(\xi,\omega,\epsilon)} q_N(t)} \leq \frac{2\omega(1 + \beta_0(\theta_j)[G^j(\tilde{\xi}) - G^j(\xi)])}{\beta_0(\theta_j)[G^j(\tilde{\xi}) - G^j(\xi)]\eta/3}.$$

This implies the claim. □

Finally, we use these claims to prove the proposition.

For any δ and ρ, there are ξ^*, $\epsilon^* > 0$ and N^* such that for all $\xi < \xi^*$, $\epsilon < \epsilon^*$, $N > N^*(\epsilon, \xi)$, and $\omega > 0$, any t' is in either $B_N(x, \Delta; \epsilon)$ for some Δ and $x > \xi + \omega$, or in $b_N(\xi, \omega, \epsilon)$ if it is at most a distance δ away from some $t \in \mathcal{T}$ with $|u_{\sigma^*}(t) - u_{\sigma_N^*}(t')| \leq \rho$.

In particular, find l such that $\theta_l \leq \bar{u} < \theta_{l+1}$, and then for any K we can construct the collection of sets

$$\left\{ B_N \left(\xi + \omega + k \frac{\theta_l - (\xi + \omega)}{K}, \frac{\theta_l - (\xi + \omega)}{K}; \epsilon \right) \right\}_{k=0}^{K-1}$$

and

$$\left\{ B_N \left(\theta_{j-1} + \eta + k \frac{\theta_j - \eta - (\theta_{j-1} + \eta)}{K}, \frac{\theta_j - \eta - (\theta_{j-1} + \eta)}{K}; \epsilon \right) \right\}_{k=0}^{K}, \quad \text{for all } j > l,$$

essentially partitioning the imitated senders by the payoffs they receive, into intervals that are disjoint, cover all attained payoffs except $[0, \xi + \omega]$ and the intervals $[\theta_j - \eta, \theta_j + \eta)$ and are arbitrarily small as $K \to \infty$.

Let $C(N, K, \xi, \omega, \eta; \epsilon)$ be the collection that is the union of these collections, and also includes, if condition 1 holds for η, the set $b_N(\xi, \omega, \epsilon)$. Call the elements of $C(N, K, \xi, \omega, \eta; \epsilon)$ by $C_1(N, K, \xi, \omega, \eta; \epsilon), \ldots, C_I(N, K, \xi, \omega, \eta; \epsilon)$.

Likewise, we can construct the collection of sets

$$\left\{ A_N \left(\xi + \omega + k \frac{\theta_l - (\xi + \omega)}{K}, \frac{\theta_l - (\xi + \omega)}{K}; \epsilon \right) \right\}_{k=0}^{K-1}$$

and

$$\left\{ A_N \left(\theta_{j-1} + \eta + k \frac{\theta_j - \eta - (\theta_{j-1} + \eta)}{K}, \frac{\theta_j - \eta - (\theta_{j-1} + \eta)}{K}; \epsilon \right) \right\}_{k=0}^{K}, \quad \text{for all } j > l,$$

which are corresponding sets of imitating types; let $D(N, K, \xi, \omega, \eta; \epsilon)$ be the collection containing these as well as $a_N(\xi, \omega, \epsilon)$ if condition 1 holds for η. Call the elements of $D(N, K, \xi, \omega, \eta; \epsilon)$ by $D_1(N, K, \xi, \omega, \eta; \epsilon), \ldots, D_I(N, K, \xi, \omega, \eta; \epsilon)$.

The proposition follows from proving that the ex-ante probability that the sender imitates some t' that is in an element of $C(N, K, \xi, \omega, \eta; \epsilon)$ converges to 1 in the large N limit and in the limit as $K \to \infty$ and $\epsilon, \omega, \xi, \eta \to 0$.

To see this, first observe that, from the above two claims, if we define

$$LB(N, K, \xi, \omega, \eta; \epsilon) = \min_i Pr(m_{\sigma_N^*}(t) \in C_i(N, K, \xi, \omega, \eta; \epsilon) | t \in D_i(N, K, \xi, \omega, \eta; \epsilon)),$$

then $\lim_{\xi \to 0} \lim_{\omega \to 0} \lim_{K \to \infty} \lim_{\epsilon \to 0} \lim_{N \to \infty} UB(N, K, \xi, \omega, \eta; \epsilon) = 1$; note that this is a uniform bound over all i.

Then, letting T denote a set of types that is an element of $C(N, K, \xi, \omega, \eta; \epsilon)$, we have

$$
\begin{aligned}
&Pr\left(m_{\sigma_N^*}(t) \in [\bigcup_{C_N(K,\xi,\omega,\eta;\epsilon)} T] \middle| \min_k |u_{\sigma_N^*}(t) - \theta_k| > \eta\right) \\
&\geq \sum_{i=1}^{I} \Bigg[Pr\left(t \in D_i(N, K, \xi, \omega, \eta; \epsilon) | \min_k |u_{\sigma_N^*}(t) - \theta_k| > \eta\right) \\
&\qquad \cdot Pr(m_{\sigma_N^*}(t) \in C_i(N, K, \xi, \omega, \eta; \epsilon) | t \in D_i(N, K, \xi, \omega, \eta; \epsilon)) \\
&\quad + Pr(m_{\sigma_N^*}(t) \in C_i(N, K, \xi, \omega, \eta; \epsilon) \Big| \min_k |u_{\sigma_N^*}(t) - \theta_k| > \eta) \Bigg] \\
&\geq \frac{Pr(t \in \bigcup_i D_i(N, K, \xi, \omega, \eta; \epsilon)) LB(N, K, \xi, \omega, \eta; \epsilon) + Pr(t \in \bigcup_i C_i(N, K, \xi, \omega, \eta; \epsilon))}{Pr(\min_k |u_{\sigma_N^*}(t) - \theta_k| > \eta)}.
\end{aligned}
\tag{A.16}
$$

Since $\lim_{\xi \to 0} \lim_{\omega \to 0} \lim_{K \to \infty} \lim_{\epsilon \to 0} \lim_{N \to \infty} PR(t \in \bigcup_i [D_i(N, K, \xi, \omega, \eta; \epsilon) \bigcup D_i(N, K, \xi, \omega, \eta; \epsilon)]) = Pr(\min_k |u_{\sigma_N^*}(t) - \theta_k > \eta)$, we have

$$
\begin{aligned}
&\lim_{\xi \to 0} \lim_{\omega \to 0} \lim_{K \to \infty} \lim_{\epsilon \to 0} \lim_{N \to \infty} \\
&\quad \frac{Pr(t \in \bigcup_i D_i(N, K, \xi, \omega, \eta; \epsilon)) LB(N, K, \xi, \omega, \eta; \epsilon) + Pr(t \in \bigcup_i C_i(N, K, \xi, \omega, \eta; \epsilon))}{Pr(\min_k |u_{\sigma_N^*}(t) - \theta_k| > \eta)} \\
&= 1
\end{aligned}
\tag{A.17}
$$

for all η, thus proving the proposition.

□

Appendix B

for "Model (non)-disclosure in supervisory stress tests"

B.0.1 Proofs of propositions

Lemma. *Assume two random variables v, w are given such that $v|w$ satisfies MLRP; that is, if $w_1 > w_2$, $f(v|w_1)/f(v|w_2)$ is increasing in v. Then $w|v$ satisfies MLRP too.*

(In words: if I get to see v first, upon seeing higher v, I expect a higher w. Then, if I get to see w first, upon seeing higher w, I expect a higher w.)

Proof. For any $v_1 > v_2$ and $w_1 > w_2$, by $v|w$ MLRP, we have $f(v_1|w_1)/f(v_1|w_2) > f(v_2|w_1)/f(v_2|w_2)$. Rearrange to obtain $f(v_1|w_1)/f(v_2|w_1) > f(v_1|w_2)/f(v_2|w_2)$

Then by Bayes rule (we use p to denote unconditional distributions), we have

$$\begin{aligned}\frac{f(w_1|v_1)}{f(w_1|v_2)} &= \frac{f(v_1|w_1)p(w_1)/p(v_1)}{f(v_2|w_1)p(w_1)/p(v_2)} \\ &= \frac{f(v_1|w_1)}{f(v_2|w_1)}\frac{p(v_2)}{p(v_1)} \\ &> \frac{f(v_1|w_2)}{f(v_2|w_2)}\frac{p(v_2)}{p(v_1)} \\ &= \frac{f(v_1|w_2)p(w_2)/p(v_1)}{f(v_2|w_2)p(w_2)/p(v_2)} = \frac{f(w_2|v_1)}{f(w_2|v_2)}\end{aligned}$$

Since this holds for any $v_1 > v_2$ and $w_1 > w_2$, $w|v$ satisfies MLRP. □

Proposition 9. *Suppose the regulator can choose any information structure to partially disclose its signal. Then the regulator can design a disclosure policy with infinitesimal noise to approximate the "efficient investment" decision that incorporates both the information*

known by the regulator and the bank. Formally, for any $\epsilon > 0$, the Fed can choose the stress test (s^*, c) and a disclosure policy π such that the bank, after observing m will invest in the risky asset if and only if $s^B \geq s^{B*}(m, c)$, and this action satisfies

$$U^{PD} = \mathbb{E}[u^F(\omega) 1_{s^B \geq s^{B*}}] > U^{FB} - \epsilon$$

where U^{FB} is the first-best utility of the regulator, and the expectation on the left-hand side is the expected regulator's utility (social welfare) under the partial disclosure policy, taken over all signal and message realizations.

Proof. Fix an $\epsilon > 0$; we claim that there exists a partial disclosure policy achieving utility $> U^{FB} - \epsilon$. We proceed in the following steps.

Step 1. We approximate the signal space into a discrete space; with a close enough approximation with large number of elements N, the distributions should converge to the distribution in continuum, and the difference in ex-ante utility from the discrete signal space and continuous signal space should differ by an infinitesimal amount.

Specifically, we assume that discrete signals $d^F, d^B|\omega$ are drawn from $\{s_1, s_2, \cdots, s_{2N}\} \subset S$, such that $P(d^F = s_i|\omega) = P(s^F \leq s_i|\omega) - P(s^F \leq s_{i-1}|\omega)$, and the endpoints are $P(d^F = s_1|\omega) = P(s^F \leq s_1)$, $P(d^F = s_i|\omega) = 1 - P(s^F|\omega)$, and analogously for $s^{B\prime}$. This should naturally define the posterior distributions $d^F|d^B$ and $\omega|d^B$. Define $s_0 = -\infty$ and $s_{2N+1} = +\infty$.

Moreover, we choose $s_1, \cdots, s_n$ such that the ex-ante probabilities $\max(P(d^F \leq s_n), P(d^B \leq s_n) < \delta$ for some δ sufficiently small: the first n signals in the discrete space happen with small probability. These correspond to 'fail' signals. We can choose $s_{N+1}, \cdots, s_{2N}$ arbitrarily, as long as they converge to the continuous distribution as $N \to \infty$. Let them be distributed of equal probability.

Step 2. Define the "optimal bank threshold" under $d^F = s_{N+i}$ for each i.

Specifically, for each $i \in \{1, 2, \cdots, n\}$, given $d^F = s_{N+i}$, choose the smallest j for which $E[u^F(\omega)|d^F = s_i, d^B = s_j] \geq 0$. Then by monotonicity, this s_j is the smallest bank's signal for which the regulator would like the bank to invest. Denote this $s_j = s^{B*}(s_{N+i})$: then by definition, bank investing iff $d^B \geq s^{B*}(d^F)$ is the first-best investment rule.

Step 3. We define the disclosure structure $\pi : m|d^F$ and the punishment threshold s^* and c as follows:

1. Define the threshold s^* to be s_n: banks fail the test and are punished iff $d^F \leq s_N$. This would guarantee ex-ante that only δ of the banks are punished.
2. Define the disclosure policy π as follows:
    - Upon observing $d^F = s_i$ $(i \geq N + 1)$, send message $m = s_i$: reveal d^F fully.

- Upon observing $d^F = s_i$ $(i < N)$, send message $m = s_i$ with probability $1 - \epsilon_i$ and send $m = s_{N+i}$ with probability ϵ_i, where $\max[\epsilon_i] < \epsilon$.

From this we immediately see that $m = d^F$ with ex-ante probability $\geq 1 - \epsilon$.

3. For each i, choose c sufficiently large and $\epsilon_i < \epsilon$ such that bank with signal $d^B = s^{B*}(s_{N+i})$ and seeing message $m = s_{N+i}$ gets expected utility ϵ' from investing in risky asset; banks will never invest if seeing signal $m \leq s_N$.

The equation defining c, ϵ_i is

$$\begin{aligned}\epsilon' =& \mathbb{E}[u^B(\omega)|d^B = s^{B*}(s_{N+i}), m = s_{N+i}] - cP[d^F \leq s_N|d^B = s^{B*}(s_{N+i}), m = s_{N+i}] \\ =& P[d^F = s_{N+i}|d^B, m = s_{N+i}]\mathbb{E}[u^B(\omega)|d^B = s^{B*}(s_{N+i}), d^F = s_{N+i}] \\ &+ P[d^F = s_i|d^B = s^{B*}(s_{N+i}), m = s_{N+i}]\mathbb{E}[u^B(\omega)|d^B = s^{B*}(s_{N+i}), d^F = s_i] \\ &- cP[d^F = s_i|d^B = s^{B*}(s_{N+i}), m = s_{N+i}]\end{aligned}$$

Write $q(\epsilon_i) = \mathbb{P}[d^F = s_i|d^B = s^{B*}(s_{N+i}), m = s_{N+i}]$. This is the posterior probability that a bank fails the test when it receives message $m = s_{N+i}$ and has signal $s^{B*}(s_{N+i})$. Also write $u_{i1} = \mathbb{E}[u^B(\omega)|d^B = s^{B*}(s_{N+i}), d^F = s_{N+i}]$ and $u_{i2} = \mathbb{E}[u^B(\omega)|d^B = s^{B*}(s_{N+i}), d^F = s_i]$. Then the above equation is simplified to:

$$\epsilon' = (1 - q(\epsilon_i))u_{i1} + q(\epsilon_i)u_{i2} - q(\epsilon_i)c$$

Now $q(\epsilon_i)$ is a continuous function on ϵ_i and $q(\epsilon_i) \to 0$ as $\epsilon_i \to 0$. So for sufficiently large c, there is some $q(\epsilon_i) \ll 1$ that satisfy this equation, and we can set $\max \epsilon_i < \epsilon$ if c is large enough.

Send $\epsilon' \to 0$ first, the bank's investment rule is: upon seeing $m = s_{N+i}$, invest if and only if $d^B \geq s^{B*}(m) = s^{B*}(s_{N+i})$. Banks will be punished with probability δ; and $\bar{d}$ and m are identical with probability at least $1 - \epsilon$.

As $\epsilon \to 0$ (and $c \to \infty$), $m = d^F$ with probability almost 1, and the bank's investment rule $d^B \geq s^{B*}(m)$ is identical to the first-best rule.

□

Appendix C

for "Nondisclosure with conflicting motives"

Proof of Theorem 10. First, since $u_s(\theta, x, a_r)$ and $u_r(\theta, a_r)$ are continuous in all arguments, as is $\tilde{h}(\theta|m)$, the expected utilities $u_s(\tilde{h}(\theta|m), x, a_r)$ and $u_r(\tilde{h}(\theta|m), a_r)$ are also continuous in all arguments. Since they are single-peaked, the peaks $a^*_{r,s}(\tilde{h}(\theta|m), x)$ and $a^*_{r,r}(\tilde{h}(\theta|m))$ are continuous in m and x as well.

Lemma 21. *Suppose that $b(m|\emptyset)$ is supported on a subset of $[c, d]$. Then $a^*_{r,r}(\beta(\theta|c)) \leq a^*_{r,r}(\beta(\theta|\emptyset)) \leq a^*_{r,r}(\beta(\theta|d))$, with strict inequality if $b(m|\emptyset)$ is not a point mass.*

Proof. Consider the derivative of the receiver's expected utility with respect to their action under message $\emptyset$ given that their belief is $b(m|\emptyset)$. At $a^*_{r,r}(\beta(\theta|\emptyset)) = a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$, the derivative should be 0. But, for any $c' < c$ or $d' > d$,

$$\frac{\partial}{\partial a_r} u_r(\beta(\theta|\emptyset), a_r)\big|_{a^*_{r,r}(\tilde{h}(\theta|c'))} = \int_m \frac{\partial}{\partial a_r} u_r(\tilde{h}(\theta|m), a_r)|_{a^*_{r,r}(\tilde{h}(\theta|c'))} b(m|\emptyset) dm > 0,$$
$$\frac{\partial}{\partial a_r} u_r(\beta(\theta|\emptyset), a_r)\big|_{a^*_{r,r}(\tilde{h}(\theta|d'))} = \int_m \frac{\partial}{\partial a_r} u_r(\tilde{h}(\theta|m), a_r)|_{a^*_{r,r}(\tilde{h}(\theta|d'))} b(m|\emptyset) dm < 0,$$

and this is true even for $c' = c$ and $d' = d$ if $\beta(\theta|\emptyset)$ is not a point distribution. □ □

Fix a putative posterior, $\beta(\theta|\emptyset)$, for the receiver given the empty message. Observe that there is an posterior on m given $\emptyset$, which can be denoted $b(m|\emptyset)$, with support on a subset of $[\underline{m}, \bar{m}]$, and

$$\beta(\theta|\emptyset) = \int_m \tilde{h}(\theta|m) b(m|\emptyset) dm.$$

Then $a^*_{r,r}(\beta(\theta|\emptyset)) \in [a^*_{r,r}(\tilde{h}(\theta|\underline{m})), a^*_{r,r}(\tilde{h}(\theta|\bar{m}))]$ by Lemma 21. The crucial step in the proof is to observe that, by intermediate value theorem, there exists $\hat{m} \in [\underline{m}, \bar{m}]$ such that $a^*_{r,r}(\beta(\theta|\emptyset)) = a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$.

The following steps show that in equilibrium $\hat{m}$ must induce senders to pool via nondisclosure.

1. Fixing $\hat{m}$, define the sets

$$\bar{A}(\hat{m}) := \{(x, m) : a^*_{r,s}(\tilde{h}(\theta|m), x) > a^*_{r,r}(\tilde{h}(\theta|\hat{m}))\},$$

$$\underline{A}(\hat{m}) := \{(x, m) : a^*_{r,s}(\tilde{h}(\theta|m), x) < a^*_{r,r}(\tilde{h}(\theta|\hat{m})).$$

2. Then define

$$\bar{B}(\hat{m}) := \bar{A}(\hat{m})) \bigcap \{(x, m) : m < \hat{m}\},$$

$$\underline{B}(\hat{m}) := \underline{A}(\hat{m})) \bigcap \{(x, m) : m > \hat{m}\}.$$

   Both are the intersection of open sets, thus also open.

3. If $(x, m) \in \bar{B}(\hat{m}) \bigcup \underline{B}(\hat{m})$, then by single-peakedness the sender prefers $a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$ to $a^*_{r,r}(\tilde{h}(\theta|m))$, and would like to withhold m instead of disclosing it.

4. Whenever $\bar{B}(\hat{m}) \bigcup \bar{B}(\underline{m})$ is nonempty, $\{m : (x, m) \in \bar{B}(\hat{m}) \bigcup \underline{B}(\hat{m})\}$ must therefore be in the support of $b(m|\emptyset)$, and $b(m|\emptyset)$ has positive measure over it.

5. Because $g(x)$ is not a point distribution, there always exists either x such that $a^*_{r,s}(\tilde{h}(\theta|\hat{m}), x) > a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$ or x such that $a^*_{r,s}(\tilde{h}(\theta|\hat{m}), x) < a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$.

   If $\hat{m} \in (\underline{m}, \bar{m})$, then the former implies that $\bar{B}(\hat{m})$ is nonempty, and the latter implies $\underline{B}(\hat{m})$ is nonempty.

6. Alternatively, if $\hat{m} = \bar{m}$ or $\hat{m} = \underline{m}$, then BSB implies that $\bar{B}(\hat{m})$ or $\underline{B}(\hat{m})$ are nonempty, respectively.

This suffices to show that all equilibria must feature pooling: there is a positive measure of signals which senders, depending on their type, have a positive probability of withholding.

As an addendum to point 6, observe that if the receiver's belief is represented by $\hat{m} = \bar{m}$, then the sender's BR induces her to withhold some signals $m < \hat{m}$, but no signals $m > \hat{m}$ (since such signals do not exist), which is inconsistent with the original belief. A similar observation holds if $\hat{m} = \underline{m}$. Thus, in equilibrium it must be that $\hat{m} \in (\underline{m}, \bar{m})$. By Lemma 21, since $b(m|\emptyset)$ is not a singleton, it places positive probability on elements to either side of $\hat{m}$.

Finally, given $\hat{m}$, there is at most one sender type that satisfies $a^*_{r,s}(\tilde{h}(\theta|\hat{m}), x) = a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$. Any type x with $a^*_{r,s}(\tilde{h}(\theta|\hat{m}), x) > a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$ will withhold for some set of signals $m < \hat{m}$, and any x with $a^*_{r,s}(\tilde{h}(\theta|\hat{m}), x) < a^*_{r,r}(\tilde{h}(\theta|\hat{m}))$ will withhold for some $m > \hat{m}$. □ □

Proof of Claim 12. Consider the case in which $a^*_{r,r}(\tilde{h}(m)) < a^*_{r,s}(\tilde{h}(m), \underline{x})$ for all m. Whenever $\hat{m} < m$, the sender prefers to send m, because $a^*_{r,r}(\tilde{h}(\hat{m})) < a^*_{r,r}(\tilde{h}(m)) < a^*_{r,s}(\tilde{h}(m), \underline{x})$. However, whenever $\hat{m} \neq \underline{m}$, for every type x there exists a positive measure of $m < \hat{m}$ such that $a^*_{r,r}(\tilde{h}(m)) < a^*_{r,r}(\tilde{h}(\hat{m})) < a^*_{r,s}(\tilde{h}(m), \underline{x})$. Therefore, if $\hat{m} \neq \underline{m}$, all withholding occurs for $m < \hat{m}$, which gives a contradiction.

When $\hat{m} = \underline{m}$, once again if $m > \hat{m}$, no sender wishes to withhold. Thus, senders may only withhold if $m = \hat{m}$, which is consistent with $\hat{m}$ representing the receiver's posterior, and results in full separation of signals. □

Proof of Claim 13. This claim follows directly from the fact that, under MBM, whenever the sender withholds, he withholds an interval of signals to one side or the other of $\hat{m}$. The sender cannot pool with other sender types, so if the sender attempts to withhold a nonempty interval of signals, the receiver should, as part of their best response, nontrivially update $\hat{m}$. Thus, $\hat{m}$ represents a fixed point of the two players' strategies only if the sender never withholds, which may be the case either when $\hat{m} \in \{\underline{m}, \bar{m}\}$ or when $m^=(\hat{m}, x) = \hat{m}$.[1] □

Proof of Prop. 12. Since utilities are single-peaked, the sender weakly prefers to withhold whenever $\hat{m}$ is inside $[m^=(m, x), m]$ or $[m, m^=(m, x)]$.

Fix x. The breakeven message $m^=(m, x)$ is strictly increasing in m, so there is a single signal M at which $m^=(x, M) = \hat{m}$. For all $m < M$, $m^=(m, x) < \hat{m}$; for all $m > M$, $m^=(m, x) > \hat{m}$. Thus, either $M < \hat{m}$, and $\hat{m} \in [m, m^=(x, m)]$ iff $m \in [M, \hat{m}]$; or, $M > \hat{m}$, and then $\hat{m} \in [m^=(x, m), m]$ iff $m \in [\hat{m}, M]$. □

[1] Visually, this is to say that $\hat{m}$ must be either on one of the boundaries or at one of the nodes at which m and $m^=(m, x)$ intersect in Figure 3.4a.

Bibliography

George Akerlof. The market for lemons: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

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

S Athey. Monotone comparative statics under uncertainty. *QJE*, 117(1):187–223, 2002.

A. Banerjee and R. Somanathan. A simple model of voice. *Quarterly Journal of Economics*, 116(1), 2001.

M. Baum and Y. Zhukov. Media ownership and news coverage of international conflict. *Political Communication*, 2018.

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

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016a.

Dirk Bergemann and Stephen Morris. Information design, bayesian persuasion, and bayes correlated equilibrium. *American Economic Review*, 106(5):586–91, 2016b.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Dirk Bergemann and Martin Pesendorfer. Information structures in optimal auctions. *Journal of Economic Theory*, 137(1):580–609, 2007.

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

J. Bull and J. Watson. Statistical evidence and the problem of robust litigation. *RAND*, 2019.

A. Chakraborty and R. Harbaugh. Persuasion by cheap talk. *American Economic Review*, 100(5), 2010.

Vincent P Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50 (6):1431–1451, 1982.

Jacques Crémer and Richard P McLean. Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica*, 56(6):1247–57, 1988.

Sulagna Dasgupta, Ilia Krasikov, and Rohit Lamba. Hard information design. *Working paper*, 2022.

Alfredo Di Tillio, Marco Ottaviani, and Peter N. Sorenson. Strategic sample selection. *Econometrica*, 89(2):911–953, 2021.

Laura Doval and Jeffrey C. Ely. Sequential information design. *Econometrica*, 88(6):2575–2608, 2020.

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.

Florian Ederer, Richard Holden, and Margaret Meyer. Gaming and strategic opacity in incentive provision. *RAND Journal of Economics*, 49(4):819–854, 2018.

Peter Eso and Balázs Szentes. Optimal information disclosure in auctions and the handicap auction. *Review of Economic Studies*, 74(3):705–731, 2007.

J. Esteban and D. Ray. Inequality, lobbying, and resource allocation. *American Economic Review*, 96(1), 2006.

Vitor Farinha Luz. Surplus extraction with rich type spaces. *Journal of Economic Theory*, 148(6):2749–2762, 2013.

J. Doyne Farmer, Alissa M. Kleinnijenhuis, Til Schuermann, and Thom Wetzer. *Handbook of Financial Stress Testing*. Cambridge University Press, 2022.

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.

Michael J. Fishman and Kathleen M. Haggerty. Investment and information acquisition. *Quarterly Journal of Economics*, 105(2):427–444, 1990.

Mark J. Flannery. Transparency and model evolution in stress testing. *Federal Reserve Bank of Boston conference paper*, 2019.

A. Frankel and N. Kartik. Muddled information. *Journal of Political Economy*, 127(4), 2019.

Guillame Frechette, Alessandro Lizzeri, and Jacopo Perego. Rules and commitment in communication: An experimental analysis. *Econometrica*, 90(5), 2022.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, 1970.

M. Gentzkow and J. Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1), 2010.

German Gieczewski. Lying by omission: Verifiable communication on networks. *JMP*, 2016.

Stefan Gissler, Jeremy Oldfather, and Doriana Ruffino. Lending on hold: Regulatory uncertainty and bank lending standards. *Journal of Monetary Economics*, 81(C):89–101, 2016.

Jacob Glazer and Ariel Rubinstein. A study in the pragmatics of persuasion: A game theoretical approach. *Theoretical Economics*, pages 395–410, 2006.

Itay Goldstein and Yaron Leitner. Stress tests and information disclosure. *Journal of Economic Theory*, 177(C):34–69, 2018.

Itay Goldstein and Yaron Leitner. Stress tests disclosure: Theory, practice, and new perspectives. In *Handbook of Financial Stress Testing*. Cambridge University Press, 2022.

Itay Goldstein and Haresh Sapra. Should banks' stress test results be disclosed? an analysis of the costs and benefits. *Foundations and Trends in Finance*, 8(1):1–54, 2014.

Jerry Green and Jean-Jacques Laffont. Partially verifiable information and mechanism design. *Review of Economic Studies*, 53(3):447–456, 1986.

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

Jeanne Hagenbach, Frederic Koessler, and Eduardo Perez-Richet. Certifiable pre-play communication: Full disclosure. *Econometrica*, 82(3):1093–1131, 2014.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, Markus Mobius, and Divyarthi Mohan. Persuading with anecdotes. *NBER working paper*, 2022.

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

Emeric Henry and Marco Ottaviani. Research and the approval process: The organization of persuasion. *American Economic Review*, 109(2):911–955, 2019.

Nicolas Inostroza and Alessandro Pavan. Persuasion in global games with application to stress testing. *Working Paper*, 2020.

Shaofei Jiang. Disclosure games with large evidence spaces. *Working paper*, 2022.

Ginger Zhe Jin, Michael Luca, and Daniel Martin. Is no news (perceived as) bad news? an experimental investigation of information disclosure. *AEJ: Microeconomics*, 13(2):141–173, 2021.

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.

Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11(1):249–272, 2019.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011a.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011b.

S. Karlin and H. Rubin. The theory of decision procedures for distributions with monotone likelihood ratio. *Annals of Mathematical Statistics*, 27(2):272–299, 1956.

Anton Kolotilin, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li. Persuasion of a privately informed receiver. *Econometrica*, 85(6):1949–1964, 2017.

Daniel Krähmer. Information disclosure and full surplus extraction in mechanism design. *Journal of Economic Theory*, 187(C):105020, 2020.

Yaron Leitner and Basil Williams. Model secrecy and stress tests. *Working Paper*, 2020.

Yaron Leitner and Basil Williams. Model secrecy and stress tests. *Journal of Finance*, Forthcoming, 2022. URL https://ssrn.com/abstract=3606654.

Hao Li and Xianwen Shi. Discriminatory information disclosure. *American Economic Review*, 107(11):3363–85, 2017.

Ying Xue Li and Burkhard Schipper. Strategic reasoning in persuasion games: An experiment. *Games and Economic Behavior*, 121:329–367, 2020.

Avi Lichtig and Ran Weksler. Information transmission in voluntary disclosure games. *Working paper*, 2022.

Steven Matthews and Andrew Postlewaite. The economics of quality testing and disclosure. *Rand Journal of Economics*, 16(3):357–364, 1985.

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

Randolph McAfee and Philip Reny. Correlated information and mechanism design. *Econometrica*, 60(2):395–421, 1992.

Dimitri Migrow and Sergei Severinov. Investment and information acquisition. *AEJ: Microeconomics*, 14(3):480–529, 2022.

Paul Milgrom. *Putting Auction Theory to Work*. Cambridge University Press, 2004.

Paul Milgrom. What the seller won't tell you: Persuasion and disclosure in markets. *Journal of Economic Perspectives*, 22(2):115–131, 2008.

Paul Milgrom and Robert Weber. The value of information in a sealed-bid auction. *Journal of Mathematical Economics*, 10(1):105–114, 1982.

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

Roger Myerson. Multistage games with communication. *Econometrica*, 54(2):323–58, 1986.

Masahiro Okuno-Fujiwara, Andrew Postlewaite, and Kotaro Suzumara. Strategic information revelation. *The Review of Economic Studies*, 57(1):25–47, 1990.

Dmitry Orlov, Andrzej Skrzypacz, and Pavel Zryumov. Design of macro-prudential stress tests. 2018 Meeting Papers 913, Society for Economic Dynamics, 2018.

Elif Osun and Erkut Ozbay. Evidence games: Lying aversion and commitment. *Working paper*, 2021.

Marco Ottaviani and Andrea Prat. The value of public information in monopoly. *Econometrica*, 69(6):1673–1683, 2001.

Markus Parlasca. Time inconsistency in stress test design. *Working Paper*, 2019.

Cecelia Parlatore and Thomas Philippon. Designing stress scenarios. *Working Paper*, 2018.

J. K.-H. Quah and B. Strulovici. Comparative statics, informativeness, and the interval dominance order. *Econometrica*, 77(6):1949–1992, 2009.

J. K.-H. Quah and B. Strulovici. Aggregating the single crossing property. *Econometrica*, 80(5):2333–2348, 2012.

David Rahman. Surplus extraction on arbitrary type spaces. *Working Paper*, 2012.

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

Daniel J. Seidmann and Eyal Winter. Strategic information transmission with verifiable messages. *Econometrica*, 65(1):163–169, 1997.

Itai Sher. Credibility and determinism in a game of persuasion. *Games and Economic Behavior*, 71(2):409–419, 2011.

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.

Denis Shiskin. Evidence acquisition and voluntary disclosure. *Working paper*, 2022.

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.

Asher Wolinsky. Information transmission when the sender's preferences are uncertain. *Games and Economic Behavior*, 42(2):319–326, 2003.

Takuro Yamashita. Optimal public information disclosure by mechanism designer. TSE Working Papers 18-936, Toulouse School of Economics (TSE), 2018.

Shughang Zhu. Transparency and model evolution in stress testing. *Working Paper*, 2018.