# From Design to Disclosure

S. Nageeb Ali, Andreas Kleiner, and Kun Zhang

Presented by Mars Leung

March 23, 2025

# Motivations

*i.e. the thing I desperately want in my life*

# Motivations

- Recall hard-information revelation game and equilibrium unraveling (*à la Grossman, 1981*)
  - Key result driver: senders rather be separated than be pooled
- What if senders prefer to conceal their own types instead? (as in a principal-agent problem)

# Motivations

- Recall hard-information revelation game and equilibrium unraveling (*à la Grossman, 1981*)
  - Key result driver: senders rather be separated than be pooled
- What if senders prefer to conceal their own types instead? (as in a principal-agent problem)
- **Punchline**: with regularity conditions, any outcome obtained through information design can *essentially* be supported at equilibrium

# Motivations

- Recall hard-information revelation game and equilibrium unraveling (*à la Grossman, 1981*)
  - Key result driver: senders rather be separated than be pooled
- What if senders prefer to conceal their own types instead? (as in a principal-agent problem)
- **Punchline**: with regularity conditions, any outcome obtained through information design can *essentially* be supported at equilibrium
- 3 ways to see this paper's contribution
  - An analysis of the disclosure game with concealment motives
  - Value of commitment in information transmission (*or lack thereof*)
  - Microfoundation of information design (*à la Kamenica & Gentzkow, 2011*)

# The Model

*The disclosure game*

# Preliminaries

- Sender observes type $\theta$ and then sends a message $m$ to Receiver, who then chooses an action $a$
- Sender type $\theta$ supported in compact and convex $\Theta \subseteq \mathbb{R}^n$
  - CDF $F$ and PDF $f$
- Message space $\mathcal{M}(\theta) := \{m \in \mathcal{C} \mid \theta \in m\}$, where $\mathcal{C}$ contains all non-empty closed subsets of $\Theta$
  - hard evidence – no lying but Sender can be vague

# Preliminaries

- Sender observes type $\theta$ and then sends a message $m$ to Receiver, who then chooses an action $a$
- Sender type $\theta$ supported in compact and convex $\Theta \subseteq \mathbb{R}^n$
  - CDF $F$ and PDF $f$
- Message space $\mathcal{M}(\theta) := \{m \in \mathcal{C} \mid \theta \in m\}$, where $\mathcal{C}$ contains all non-empty closed subsets of $\Theta$
  - hard evidence – no lying but Sender can be vague
  - e.g. $m = \{\theta^*\}$ – "my type is precisely $\theta^*$"
  - e.g. $m = [\underline{\theta}, \bar{\theta}]$ – "my type is somewhere between $\underline{\theta}$ and $\bar{\theta}$
  - e.g. $m = \Theta$ –"I'm not telling you anything"

# Preliminaries

- Sender observes type $\theta$ and then sends a message $m$ to Receiver, who then chooses an action $a$
- Sender type $\theta$ supported in compact and convex $\Theta \subseteq \mathbb{R}^n$
  - CDF $F$ and PDF $f$
- Message space $\mathcal{M}(\theta) := \{m \in \mathcal{C} \mid \theta \in m\}$, where $\mathcal{C}$ contains all non-empty closed subsets of $\Theta$
  - hard evidence – no lying but Sender can be vague
  - e.g. $m = \{\theta^*\}$ – "my type is precisely $\theta^*$"
  - e.g. $m = [\underline{\theta}, \bar{\theta}]$ – "my type is somewhere between $\underline{\theta}$ and $\bar{\theta}$
  - e.g. $m = \Theta$ –"I'm not telling you anything"
- Action space $A$ compact and metrizable
- Sender's payoff $u_S(a, \theta)$ continuous in $a$ for each $\theta$
- Receiver's payoff $u_R(a, \theta)$ upper-semicontinuous in $a$ for each $\theta$

# Strategies, Beliefs, and Equilibria

- Sender's strategy $\rho : \Theta \to \Delta(\mathcal{C})$, where $\forall \theta \in \Theta : \rho(\theta) \in \Delta\big(\mathcal{M}(\theta)\big)$
- Receiver's strategy $\tau : \mathcal{C} \to A$
- Receiver's belief $\mu : \mathcal{C} \to \Delta(\theta)$

## Definition (Perfect Bayesian Equilibrium)

An assessment $(\rho, \tau, \mu)$ is a PBE if it satisfies:

1. Sequential rationality:

$$\begin{cases} \forall \theta, m: & \rho(m|\theta) > 0 \iff m \in \operatorname{argmax}_{m \in \mathcal{M}(\theta)} u_S(\tau(m), \theta) \\ \forall m: & \tau(m) \in \operatorname{argmax}_{a \in A} \int_\Theta u_R(a, \theta) d\mu(\theta|m) \end{cases}$$

2. Consistency: $\mu$ is obtained on-path from $F$ given $\rho$ using Bayes' rule
3. Evidence respect:

$$\forall \theta, m: \quad \mu(\theta|m) > 0 \text{ only if } \theta \in m$$

## The Main Result

*3 more assumptions plus 1 benchmark before the punchline*

# The Key Behavioural Assumption

- Let the Receiver's best-response correspondence given belief $G$ be $a^*(G) := \text{argmax}_{a \in A} \int u_R(a, \theta) dG(\theta)$
- Abusing notation:
  if $G$ corresponds to full knowledge of a type $\theta$, then write $a^*(\theta)$

**Assumption 1 (Concealment Motive)**

For every type $\theta$, belief $G$ such that $\theta \in \text{supp}(G)$,

$$\forall a \in a^*(G): \quad u_S(a, \theta) \geq u_S\big(a^*(\theta), \theta\big)$$

- Sender never benefits from full revelation

# Off-path Regulation

**Assumption 2 (Worst-case Type)**

Every message $m$ contains a worst-case type $\hat{\theta}_m$ such that:

$$\forall \theta \in m: \quad u_S\big(a^*(\theta), \theta\big) \geq u_S(a^*(\hat{\theta}_m), \theta\big)$$

- *"I would rather reveal my type than be known as a $\hat{\theta}_m$"*
- To deter off-path messages when constructing equilibria

# Continuity

- Let the Receiver's expected payoff with belief $G$ be $U_R(G) := \max_{a \in A} \int u_R(a, \theta) dG(\theta)$
- Given belief $G$, let the best-response actions that maximise/minimise Sender's payoffs in expectation be:

$$\begin{cases} \bar{a}(G) & \in \operatorname{argmax}_{a \in a^*(G)} \int u_S(a, \theta) dG(\theta) \\ \underline{a}(G) & \in \operatorname{argmin}_{a \in a^*(G)} \int u_S(a, \theta) dG(\theta) \end{cases}$$

## Assumption 3 (Continuity)

- $U_R(G)$ is continuous and $a^*(G)$ is upper hemi-continuous;
- For every belief $G$, $\varepsilon > 0$ and $\delta > 0$, there exist:
  - a belief $\bar{H}$ $\varepsilon$-close to $G$ s.t. any best response to $\bar{H}$ is $\delta$-close to $\bar{a}(G)$;
  - a belief $\underline{H}$ $\varepsilon$-close to $G$ s.t. any best response to $\underline{H}$ is $\delta$-close to $\underline{a}(G)$;
- The functions sending $G$ to $\bar{H}$ and $\underline{H}$ are both measurable

# The Information-design Benchmark

- Sender commits publicly to a Blackwell experiment before $\theta$ is realised
- Receiver observes the experiment's realisation and takes an action
- Receiver's posterior belief $G \in \Delta\Theta$ is distributed according to $\sigma$ which satisfies Bayes plausibility (BP): $\int G d\sigma(G) = F$
- Let's call a payoff profile $(u_S^*, u_R^*)$

# The Information-design Benchmark

- Sender commits publicly to a Blackwell experiment before $\theta$ is realised
- Receiver observes the experiment's realisation and takes an action
- Receiver's posterior belief $G \in \Delta\Theta$ is distributed according to $\sigma$ which satisfies Bayes plausibility (BP): $\int G d\sigma(G) = F$
- Let's call a payoff profile $(u_S^*, u_R^*)$ *achievable* if we can find such a Bayes-plausible distribution of posteriors that induces it in ex ante expectation

# The Punchline

**Theorem 1**

Under Assumptions 1-3, for every achievable payoff profile $(u_S^*, u_R^*)$ and $\varepsilon > 0$, there is a disclosure-game equilibrium whose payoffs are $\varepsilon$-close

# The Punchline

**Theorem 1**

Under Assumptions 1-3, for every achievable payoff profile $(u_S^*, u_R^*)$ and $\varepsilon > 0$, there is a disclosure-game equilibrium whose payoffs are $\varepsilon$-close

- If an outcome is achievable by information design, then I can guarantee you that there is an equilibrium in the disclosure game that *essentially* sustains the same outcome
  - The reverse is trivially true
  - Achievability by information design $\overset{\sim}{\iff}$ supportability by disclosure-game equilibrium under A1-3
- Stark contrast with classical unravelling results
- Sender does not accrue meaningful *(or any)* gain from the commitment power of an information designer

# The Finite Analogue

- Theorem 2 extends the result to $F$ with finite support
- As long as no single type has a probability mass too large, the result goes through

## Theorem 2

Under Assumptions 1-3, for every $\varepsilon > 0$, there is $\gamma > 0$, such that:
if $F$ has finite support with $F(\{\theta\}) \leq \gamma$ for every type $\theta$,
then for every achievable payoff profile $(u_S^*, u_R^*)$,
there exists a disclosure-game equilibrium whose payoffs are $\varepsilon$-close

# Proof of Theorem 1

*A lot of measure-theory magic*

# Terminology

- A distribution over $\Theta$: a *belief*
- A distribution of beliefs that satisfies (BP): $\sigma$ – a *segmentation*
- A member of a segmentation's support: $G \in \text{supp}(\sigma)$ – a *segment*
- A payoff profile is *achieved* by a segmentation if it arises as ex ante expected payoffs given best responses
- A payoff profile is *achievable* if there exists a segmentation that achieves said payoffs in ex ante expectation given best responses

# Proof Outline

**Definition (Finite Segmentation)**

A segmentation $\sigma$ is **finite** if its support is a finite set

**Definition (Partitional Segmentation)**

A segmentation $\sigma$ is **partitional** if for every $G, H \in \text{supp}(\sigma)$ s.t. $G \neq H$, $\text{supp}(G) \cap \text{supp}(H)$ has zero $F$-measure

- Lemma 1 proves that any payoff profile achieved by a finite and partitional segmentation is supportable at equilibrium
- Lemma 2 proves that any payoff profile achievable can be approximately achieved by a finite and partitional segmentation

# Lemma 1 – Proof

## Lemma 1

Under Assumptions 1-2, every payoff profile achieved by a finite partitional segmentation can be supported as an equilibrium

# Lemma 1 – Proof

**Lemma 1**

Under Assumptions 1-2, every payoff profile achieved by a finite partitional segmentation can be supported as an equilibrium

- Given an arbitrary finite partitional segmentation $\sigma$ and a corresponding best response $a : \text{supp}(\sigma) \to A$
- On path:
  - Sender first finds a segment $G \in \text{supp}(\sigma)$ where $\theta \in \text{supp}(G)$
  - Almost surely, there is only one such $G$
  - Sender sends $m = \text{supp(G)}$
  - Receiver then updates to belief $G$ and plays $a(G)$
- Off path:
  - Receiver punishes by playing the best response to the worst case $a(\hat{\theta}_m)$
  - It exists by Assumption 2
- Easy to verify sequentially rationality and belief consistency
  - No profitable deviations by Assumptions 1-2 □

# Lemma 2 – Proof 1/2

## Lemma 2

Under Assumption 3, for every achievable payoff profile $(u_S^*, u_R^*)$ and every $\varepsilon > 0$, there exists a finite partitional segmentation that achieves payoffs within $\varepsilon$ of $(u_S^*, u_R^*)$

- Given an arbitrary $\varepsilon > 0$ and payoff profile $(u_S^*, u_R^*)$, along with its corresponding segmentation $\sigma$ and best response

# Lemma 2 – Proof 1/2

## Lemma 2

Under Assumption 3, for every achievable payoff profile $(u_S^*, u_R^*)$ and every $\varepsilon > 0$, there exists a finite partitional segmentation that achieves payoffs within $\varepsilon$ of $(u_S^*, u_R^*)$

- Given an arbitrary $\varepsilon > 0$ and payoff profile $(u_S^*, u_R^*)$, along with its corresponding segmentation $\sigma$ and best response
- Assume for now the Receiver specifically plays $\bar{a}(G)$ given any $G$
  - The proof goes through analogously for $\underline{a}(G)$
  - Anything in between is a convex combination between the two cases, supportable by Receiver playing mixed strategies

# Lemma 2 – Proof 1/2

## Lemma 2

Under Assumption 3, for every achievable payoff profile $(u_S^*, u_R^*)$ and every $\varepsilon > 0$, there exists a finite partitional segmentation that achieves payoffs within $\varepsilon$ of $(u_S^*, u_R^*)$

- Given an arbitrary $\varepsilon > 0$ and payoff profile $(u_S^*, u_R^*)$, along with its corresponding segmentation $\sigma$ and best response
- Assume for now the Receiver specifically plays $\bar{a}(G)$ given any $G$
  - The proof goes through analogously for $\underline{a}(G)$
  - Anything in between is a convex combination between the two cases, supportable by Receiver playing mixed strategies
- 3-step procedures to approximate $\sigma$:

$$\underset{\text{given}}{\overset{\sigma}{}} \longrightarrow \underset{\text{best responses close to } \bar{a}}{\overset{\sigma_1}{}} \longrightarrow \underset{\text{finite}}{\overset{\sigma_2}{}} \longrightarrow \underset{\text{finite \& partitional}}{\overset{\sigma_3}{}}$$

# Lemma 2 – Proof 2/2

1. Construct $\sigma_1$: replace each segment $G \in \sigma$ with a nearby segment $\bar{H}$ such that any best response to $\bar{H}$ is arbitrarily close to $\bar{a}(G)$
   - We can do this by direct assumption

2. Construct $\sigma_2$: merge segments in $\sigma_1$ that are close to each other into a single segment until we have only finitely many segments
   - For any $G \in \text{supp}(\sigma_1)$, there is a small open ball around $G$ s.t. the best response does not vary much, by best-response continuity
   - We can cover the all of $\text{supp}(G)$ with finitely many balls and attach an aggregated segment to each ball

3. Construct $\sigma_3$: partition the type space very fine and approximate each $G \in \sigma_2$ using a collection of partition cells
   - We can do this because $F$ is absolutely continuous $\square$

# Proof Outline

**Theorem 1**

Under Assumptions 1-3, for every achievable payoff profile $(u_S^*, u_R^*)$ and $\varepsilon > 0$, there is a disclosure-game equilibrium whose payoffs are $\varepsilon$-close

- Lemma 1 proves that any payoff profile achieved by a finite and partitional segmentation is supportable at equilibrium
- Lemma 2 proves that any payoff profile achievable can be approximately achieved by a finite and partitional segmentation
- Theorem 1 proven by applying Lemmas 2 then 1 ■

# Counterexamples and Examples

*With a splash of philosophy at the end*

# Failure of Assumption 1: Motive to Separate

**Assumption 1 (Concealment Motive)**

For every type $\theta$, belief $G$ such that $\theta \in \text{supp}(G)$,

$$\forall a \in a^*(G): \quad u_S(a,\theta) \geq u_S\big(a^*(\theta),\theta\big)$$

- Consider a more canonical setup of the dsiclosure game:
- $u_S(a,\theta)$ is constant in $\theta$ but increasing in $a$
- $u_R(a,\theta) = -(a-\theta)^2$
- Unique equilibrium entails unraveling

# Failure of Assumption 1: Motive to Separate

## Assumption 1 (Concealment Motive)

For every type $\theta$, belief $G$ such that $\theta \in \text{supp}(G)$,

$$\forall a \in a^*(G): \quad u_S(a,\theta) \geq u_S\big(a^*(\theta),\theta\big)$$

- Consider a more canonical setup of the dsiclosure game:
- $u_S(a,\theta)$ is constant in $\theta$ but increasing in $a$
- $u_R(a,\theta) = -(a-\theta)^2$
- Unique equilibrium entails unraveling
- However, if Sender is risk averse, the optimal payoff achievable for type-1 Sender by information design comes from no revelation
  - The resultant ex ante expected payoffs cannot be approximately supportable by any disclosure-game equilibrium

# Failure of Assumption 2: No Worst-case Type

## Assumption 2 (Worst-case Type)

Every message $m$ contains a worst-case type $\hat{\theta}_m$ such that:

$$\forall \theta \in m: \quad u_S\big(a^*(\theta), \theta\big) \geq u_S(a^*(\hat{\theta}_m), \theta\big)$$

- Say $A = \{1, 2\}$, $\theta \sim \text{Uni}[0, 1]$, and:

$$\begin{cases} u_R(a = 1, \theta) = 1 - \theta; & u_S(a = 1, \theta) = \mathbb{I}\{\theta \geq 1/2\} \\ u_R(a = 2, \theta) = \theta; & u_S(a = 2, \theta) = 1/2 \end{cases}$$

- No worst-case type $\hat{\theta}_m$ exists when $m = \Theta$

# Failure of Assumption 2: No Worst-case Type

## Assumption 2 (Worst-case Type)

Every message $m$ contains a worst-case type $\hat{\theta}_m$ such that:

$$\forall \theta \in m: \quad u_S\big(a^*(\theta), \theta\big) \geq u_S(a^*(\hat{\theta}_m), \theta\big)$$

- Say $A = \{1, 2\}$, $\theta \sim \text{Uni}[0, 1]$, and:

$$\begin{cases} u_R(a = 1, \theta) = 1 - \theta; & u_S(a = 1, \theta) = \mathbb{I}\{\theta \geq 1/2\} \\ u_R(a = 2, \theta) = \theta; & u_S(a = 2, \theta) = 1/2 \end{cases}$$

- No worst-case type $\hat{\theta}_m$ exists when $m = \Theta$
- Then, we can show full revelation is not approximately supportable by any disclosure-game equilibrium
  - There always exists a profitable deviation to the off-path strategy $m = \Theta$ since we lack a punishment that works for all types

# Failure of Assumption 3: Best-response Discontinuity

## Assumption 3 (Continuity)

...

- For every belief $G$, $\varepsilon > 0$ and $\delta > 0$, there exist:
  - a belief $\bar{H}$ $\varepsilon$-close to $G$ s.t. any best response to $\bar{H}$ is $\delta$-close to $\bar{a}(G)$;

...

- Say $A = \{0, 1, ..., n\}$ and $\Theta = \{\theta_1, ..., \theta_n\}$
- $u_S(a, \theta) = \mathbb{I}\{a = 0\}$ and $u_R(a, \theta) = \mathbb{I}\{a = 0\} + n \cdot \mathbb{I}\{\theta = \theta_a\}$
- Any $a$ is a best response under a uniform $G$, but $a = 0$ becomes sub-optimal as soon as $G$ departs ever so slightly from uniformity
- Then unless the prior is exactly uniform at equilibrium, Receiver will never pick $a = 0$ and Sender will always receive zero payoff

# Failure of Assumption 3: Best-response Discontinuity

## Assumption 3 (Continuity)

...

- For every belief $G$, $\varepsilon > 0$ and $\delta > 0$, there exist:
  - a belief $\bar{H}$ $\varepsilon$-close to $G$ s.t. any best response to $\bar{H}$ is $\delta$-close to $\bar{a}(G)$;

...

- Say $A = \{0, 1, ..., n\}$ and $\Theta = \{\theta_1, ..., \theta_n\}$
- $u_S(a, \theta) = \mathbb{I}\{a = 0\}$ and $u_R(a, \theta) = \mathbb{I}\{a = 0\} + n \cdot \mathbb{I}\{\theta = \theta_a\}$
- Any $a$ is a best response under a uniform $G$, but $a = 0$ becomes sub-optimal as soon as $G$ departs ever so slightly from uniformity
- Then unless the prior is exactly uniform at equilibrium, Receiver will never pick $a = 0$ and Sender will always receive zero payoff
- Any non-zero Sender's payoff is not supportable

# Example: Monopoly Pricing

- Consumer $S$ sends $m \in \mathcal{M}(\theta)$ to Monopolist $R$, who sets price $a \in \mathbb{R}_{\geq 0}$
- Consumer has a private valuation $\theta \sim F[\underline{\theta}, \bar{\theta}]$
  - Unit demand: $u_S(a, \theta) = \max\{\theta - a, 0\}$ and $u_R = a \cdot \mathbb{I}\{a \leq \theta\}$
- The literature gives an information-design benchmark:

**BBM Triangle (Bergemann, Brooks & Morris, 2015)**

A payoff profile is achievable as long as:

- Consumer's payoff is nonnegative;
- Monopolist is no worse off than posting the optimal uniform pice; and
- Total payoff is no more than maximal aggregate surplus

- We can verify Assumptions 1-3, and then by Theorem 1, conclude that BBM Triangle also *approximately* characterises the set of supportable equilibrium payoff in a disclosure game

# Information Design as a Metaphor

- The authors quote Bergemann and Morris (2019) regarding difficulty in interpreting information design:

  *"...giving a literal information design interpretation...is more subtle. We need to identify an information designer who knew consumers' valuations and committed to give partial information to the monopolist in order to maximize the sum of consumers' welfare. Importantly, even tough the disclosure rule is optimal for consumers as a group, individual consumers would not have an incentive to truthfully report their valuations to the information designer..."*

# Information Design as a Metaphor

- The authors quote Bergemann and Morris (2019) regarding difficulty in interpreting information design:

  *"...giving a literal information design interpretation...is more subtle. We need to identify an information designer who knew consumers' valuations and committed to give partial information to the monopolist in order to maximize the sum of consumers' welfare. Importantly, even tough the disclosure rule is optimal for consumers as a group, individual consumers would not have an incentive to truthfully report their valuations to the information designer..."*

- But an intermediary or commitment is not needed if we ground the achievable outcomes in a hard-evidence disclosure setting
- With clever equilibrium refinement e.g. truth-telling refinement (Hart, Kremer & Perry 2017), we can select the optimal outcome achievable

# Conclusion

*"I hope this paper does for information design like what Rubinstein did for Nash bargaining"*

*— me, earlier this week*

# Discussions

- Do I think it has successfully *micro-founded* information design?

# Discussions

- Do I think it has successfully *micro-founded* information design?
- Yes, but not as clean as I hoped

# Discussions

- Do I think it has successfully *micro-founded* information design?
- Yes, but not as clean as I hoped
- Assumption 3 seems hyper-specific and not at all easy to verify
- The characterisation is not sharp: assumptions are sufficient but not necessary
- Mulitplicity of equilibria makes predictions difficult
  - Additional refinements needed to select sensible equilibria

# Discussions

- Do I think it has successfully *micro-founded* information design?
- Yes, but not as clean as I hoped
- Assumption 3 seems hyper-specific and not at all easy to verify
- The characterisation is not sharp: assumptions are sufficient but not necessary
- Mulitplicity of equilibria makes predictions difficult
  - Additional refinements needed to select sensible equilibria
  - If you think about it, this paper really only micro-founded information-design *achievability*, but not information-design *solution*
  - Optimal information design as an equilibrium selection

Thank you!

*:D*