2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

On the Price of Transparency: A Comparison between Overt Persuasion and Covert Signaling

Tao Li and Quanyan Zhu

***Abstract*— Transparency of information disclosure has always been considered an instrumental component of effective governance, accountability, and ethical behavior in any organization or system. However, a natural question follows: *what is the cost or benefit of being transparent*, as one may suspect that transparency imposes additional constraints on the information structure, decreasing the maneuverability of the information provider. This work proposes and quantitatively investigates the *price of transparency* (PoT) in strategic information disclosure by comparing the perfect Bayesian equilibrium payoffs under two representative information structures: overt persuasion and covert signaling models. PoT is defined as the ratio between the payoff outcomes in covert and overt interactions. As the main contribution, this work develops a two-stage-bilinear (TSB) programming approach to solve for non-degenerate perfect Bayesian equilibria of dynamic incomplete information games with finite states and actions. Using TSB, we show that it is always in the information provider's interest to choose the transparent information structure, as $0 \leq$ PoT ≤ 1. The upper bound is attainable for any strictly Bayesian-posterior competitive games, of which zero-sum games are a particular case. For continuous games, the PoT, still upper-bounded by 1, can be arbitrarily close to 0, indicating the tightness of the lower bound. This tight lower bound suggests that the lack of transparency can result in significant loss for the provider.**

I. Introduction

Information asymmetry refers to the imbalance among decision-makers in their knowledge of relevant factors or details. The double-edged nature of the imbalance of power caused by asymmetric information is noteworthy. On the one hand, it can foster the development of deception-based defense mechanisms that benefit the cybersecurity realm [1]. On the other hand, it can also result in performance loss in adversarial machine leanring [2].

One natural remedy to this asymmetry is to increase transparency in information disclosure. However, *what is the price of being transparent* the information provider (the sender) has to pay, as transparency requirements may impose additional constraints on the sender's side? Does increased transparency lead to decreased maneuverability for the sender, thereby impairing the effectiveness of systems built on information asymmetry, such as cyber deception in security applications? Does one have to choose between ethical standards and operational effectiveness?

As information asymmetry is prevalent in security applications [3] and other real-world systems [4], investigating the price of transparency (PoT) is imperative. This work initiates a quantitative study on PoT in strategic information disclosure. We consider a communication game between the sender and the receiver, where the state of nature is privately revealed to the sender only. Possessing this informational advantage, the sender discloses partial information (signal) regarding the state to the receiver to manipulate its belief, leading the receiver to take actions favored by the sender. The information disclosure mechanism (i.e., how the sender creates the signal) is referred to as the information structure in the literature [5]. To answer the questions above on the transparency of information disclosure, we compare two information structures: 1) overt persuasion (OP), where the sender publicly announces its mechanism, creating a transparent information disclosure, and 2) covert signaling (CS), where the mechanism is kept private throughout the gameplay, and the receiver only observes the signal. The two information structures are summarized in Figure 1.

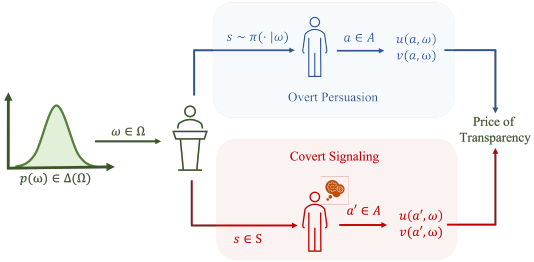

Fig. 1: A schematic illustration of two juxtaposed communication games with distinct information structures. The state variable ω is randomly generated from the prior distribution p and privately revealed to the sender. The receiver must infer the current state using the signal s from the sender and then best respond to its belief. The payoff to the sender (the receiver), denoted by $u(a,\omega)[v(a,\omega)]$, is jointly determined by the state ω and the receiver's action a. The key difference between the two information structures is whether the signaling mechanism π is made public or not.

PoT is defined as the ratio of the sender's equilibrium payoff under covert signaling over its counterpart under overt persuasion to quantify the price of choosing the transparent information structure. Note that the equilibrium concept considered here is the perfect Bayesian equilibrium (PBE), as strategic information disclosure studied in this paper is a dynamic game of incomplete information, and players are assumed sequentially rational [6].

As the transparency requirement mandates the sender to reveal its intention on signaling, it seems to give the receiver an upper hand. However, as opposed to the first impression,

Authors are with the Department of Electrical and Computer Engineering, New York University, NY, 11201, USA. tl2636, qz494@nyu.edu.
This work is partially supported by grants ECCS-1847056 and BCS-2122060 from National Science Foundation (NSF) and grant W911NF-19-1-0041 from Army Research Office (ARO). Full version available at https://arxiv.org/pdf/2304.00096.pdf

the key finding is that PoT ≤ 1 for any communication games, indicating that opting for the transparent information structure (OP) does not degrade the sender's payoff. On the contrary, the opaque one (CS) creates "friction" during the information transmission: the receiver needs to conjecture the sender's mechanism first, and then the conjecture must satisfy the consistency requirement in PBE. Consequently, covert signaling imposes more constraints on players' admissible strategies than overt persuasion. In comparison, transparent information disclosure leads to efficient communication, as players need not consider consistency.

Contributions: Our main contributions include 1) the development of a two-stage-bilinear (TSB) programming approach (Theorem 2) for solving non-degenerate PBE in strategic information disclosure; 2) the identification of a special class of communication games, termed strictly Bayesian-posterior competitive games, for which the upper bound is attained: PoT $= 1$, (Theorem 3); 3) the construction of a family of quadratic games for which PoT can be arbitrarily close to 0, indicating the tightness of the lower bound (Theorem 4).

Related Works: This work stands at the intersection of two lines of research: strategic information transmission and algorithmic information design. The two information structures are inherited from the Bayesian persuasion model in [7] and the signaling model in [8], respectively. Starting from these seminal models, this work carries out a comparative study of the two information structures. Unlike early comparative studies [9], [10] focusing on Bayesian Nash equilibrium, this work treats perfect Bayesian equilibrium, a more challenging concept involving belief consistency.

This work also subscribes to the recent line of works that explores the computational aspect of information structure design [11]–[14]. These mentioned works provide hardness results on the computational complexity of solving for the equilibrium information structures without showing concrete algorithms. In contrast, not concerning the existence of PBE or the associated complexity, we present a bilinear programming approach to compute PBE, which in turn corroborates these hardness results in [11]–[14].

II. Strategic Information Disclosure: Persuasion and Signaling

Consider a communication game as in [8], where the better-informed sender, upon receiving the state of nature, sends a signal to the receiver who then takes an action that determines payoffs to both players. Mathematically, the game model is given by a tuple $\langle \Omega, \mathcal{S}, \mathcal{A}, v, u \rangle$, where 1) Ω is the set of possible states with its typical element denoted by ω, and the realization of ω is only revealed to the sender; 2) $p \in \Delta(\Omega)$ denotes the prior distribution over the state space; 3) $\mathcal{S}$ is the set of signals possessed by the sender with its typical element denoted by s; 4) $\mathcal{A}$ is the action space of the receiver; 5) $u, v : \Omega \times \mathcal{A} \to \mathbb{R}$ are non-negative utilities of the sender and the receiver, respectively. Here, $\Delta(\cdot)$ denotes the set of all probability measures compatible with the underlying σ-algebra (e.g., Borel) over the set of interest.

Information Structures: The information structure concerns how the sender signals to the receiver. An information structure or signaling mechanism is defined by a mapping $\pi : \Omega \to \Delta(\mathcal{S})$, i.e., $\pi(\cdot|\omega)$ is a probability distribution over the signal space.

Covert Signaling (CS): As shown in Figure 1, the information structure π is unknown to the receiver who consequently cannot form a posterior belief $\lambda(\cdot|s) \in \Delta(\Omega)$, as the Bayes update requires the knowledge of the information structure π: $\lambda(\omega|s) = \frac{\pi(s|\omega)p(\omega)}{\int_\Omega \pi(s|\omega')p(d\omega')}$. In this case, the receiver can begin with a conjectural belief system, and best respond to these beliefs. If the belief system is consistent with the receiver's and the sender's strategies in the Bayesian sense, then the belief system and players' strategies constitute a perfect Bayesian equilibrium (PBE) [8].

Definition 1: A triple of the sender's information structure π, the receiver's strategy $\alpha : \mathcal{S} \to \Delta(\mathcal{A})$, and a belief system $\lambda : \mathcal{S} \to \Delta(\Omega)$ is a perfect Bayesian equilibrium if it satisfies (1): among all admissible information structures, π maximizes the sender's expected payoff given α; (2): for any signal s, $\alpha(s)$ maximizes the receiver's expected payoff under λ; (3): the belief system is consistent with π and α.

$$\pi \in \underset{\mu:\Omega\to\Delta(\mathcal{S})}{\arg\max} \int_\Omega \int_{\mathcal{S}} \int_{\mathcal{A}} u(a,\omega)\alpha(da|s)\mu(ds|\omega)p(d\omega), \quad (1)$$

$$\alpha(\cdot|s) \in \underset{\mu\in\Delta(\mathcal{A})}{\arg\max} \int_{\mathcal{A}} v(a,\omega)\mu(da)\lambda(d\omega|s), \quad (2)$$

$$\lambda(\omega|s) = \pi(s|\omega)p(\omega) \Big/ \int_\Omega \pi(s|\omega')p(d\omega'), \text{ if } \gamma(s) > 0, \quad (3)$$

where $\gamma(s) = \int_\Omega \pi(s|\omega')p(d\omega')$ denotes the probability of generating a particular signal s. We refer to $\mathrm{supp}(\gamma)$ as the set of realizable signals. The consistency (3) in PBE requires that for any realizable signal, the belief system conforms to the Bayes rule under equilibrium strategies π and α. For those unrealizable (the receiver never observes these signals), the beliefs can be arbitrary ones, as they never appear on the equilibrium path [6, Chapter 8.2], which makes no difference to the equilibrium strategies. Yet, these arbitrary beliefs may cause trouble in our bilinear programming formulation in finite games presented in Section III. Hence, we restrict the focus to the non-dengenrate cases. For a triple (π, α, λ), if it satisfies (1), (2), (3), and $\mathrm{supp}(\gamma) = \mathcal{S}$, it is a non-degenerate PBE: every signal yields an equilibrium path.

Overt Persuasion (OP): Unlike covert signaling, the sender in overt persuasion first reveals π to the receiver and then draws a signal according to $\pi(\cdot|\omega)$, when the realized state is ω. Hence, the receiver need not conjecture, as the belief $\lambda(\omega|s)$ is readily available through Bayesian update once the signal is observed. The equilibrium information structure is given by (assuming non-degenerate equilibrium)

$$\begin{aligned} \max_{\pi:\Omega\to\Delta(\mathcal{S})} & \int_\Omega \int_{\mathcal{S}} \int_{\mathcal{A}} u(a,\omega)\alpha(da|s)\pi(ds|\omega)p(d\omega) \\ \text{s.t. } & \alpha(\cdot|s) \in \underset{\mu\in\Delta(\mathcal{A})}{\arg\max} \int_{\mathcal{A}} v(a,\omega)\mu(da)\lambda(d\omega|s), \quad (4) \\ & \lambda(\omega|s) = \frac{\pi(s|\omega)p(\omega)}{\int_\Omega \pi(s|\omega')p(d\omega')}. \end{aligned}$$

Bayesian Plausibility: Since (4) is a bilevel optimization of functionals, directly solving for π remains challenging. A key observation in [7] is that an information structure is equivalent to a distribution over posterior beliefs. Recall that each signal s in OP leads to a posterior belief $\lambda(s) \in \Delta(\Omega)$ with respect to the information structure π. Accordingly, each information structure π leads to a distribution over posterior beliefs. Denote a distribution of posteriors by $\tau \in \Delta(\Delta(\Omega))$, and τ is given by $\tau(\lambda) = \int_{s:\lambda=\lambda(\cdot|s)} \int_\Omega \pi(s|\omega)p(d\omega)ds$, assuming that $\{s : \lambda = \lambda(\cdot|s)\}$ is a measurable subset of $\mathcal{S}$, and ds denotes the Borel measure. This work considers the cases where Ω is a separable metric space (e.g., $\mathbb{R}^n$ or finite sets). Consequently, $\Delta(\Omega)$, endowed with weak*-topology, is also separable and metrizable. Hence, the Borel probability measure is well-defined on $\Delta(\Omega)$.

With a slight abuse of notation, we also denote by λ an individual belief in $\Delta(\Omega)$. A belief is Bayesian inducible under π if $\tau(\lambda) > 0$, i.e., $\lambda \in \text{supp}(\tau)$, and distribution of posteriors τ is *Bayesian plausible* if the expected posterior probability equals the prior: $\int \lambda\tau(d\lambda) = p$. [7] finds that for any Bayesian plausible distribution τ, one can always find an information structure π such that every $\lambda \in \text{supp}(\tau)$ is π-Bayesian inducible. With this observation, searching for the optimal information structure is equivalent to finding the optimal posteriors distribution through backward induction specified below. Given a posterior belief λ, denote by $\hat{a}(\lambda) = \arg\max_a \mathbb{E}_{\omega\sim\lambda} v(a, \omega)$ the best response of the receiver, which is assumed to be a singleton (tie breaks in favor of the sender). Under this belief, the sender's expected utility is $\hat{u}(\lambda) = \mathbb{E}_\lambda u(\hat{a}(\lambda), \omega)$. If λ is further subject to a distribution τ, then the sender's payoff is $\mathbb{E}_\tau \hat{u}(\lambda)$. Since the sender's goal is to find the distribution τ that maximizes his expected utility, the corresponding optimization problem is given by

$$\max_\tau \mathbb{E}_\tau \hat{u}(\lambda), \quad \text{s.t.} \int \lambda\tau(d\lambda) = p. \tag{5}$$

Price of Transparency Denote by U^{CS} and U^{OP} the sender's equilibrium payoff in covert signaling and overt persuasion, respectively. The price of transparency (PoT) is defined as $\text{PoT} = U^{CS}/U^{OP}$. Note that U^{OP}, the optimal value in (4) [or equivalently (5)], is unique. In contrast, the communication game in CS may admit multiple equilibria and hence, different equilibrium payoffs. Given that PoT is not a definite number but a collection of possibilities, the rest of paper aims to identify its upper and lower bounds.

III. The Price of Transparency in Finite Games

Matrix Representation of Information Structure Our treatment of PoT begins with finite games where $\Omega, \mathcal{S}$, and $\mathcal{A}$ are all finite discrete sets. In finite games, the sender's and the receiver's strategies and the belief system all take matrix forms. We introduce some notations in the following to facilitate the discussion. Let $\Omega = \{\omega_i\}_{i\in[M]}, \mathcal{S} = \{s_i\}_{i\in[N]}$, and $\mathcal{A} = \{a_i\}_{i\in[K]}$, where $[n] := \{1, 2, \ldots, n\}, n \in \mathbb{N}_+$. Assume that $N \geq M$. Denote by $p \in \mathbb{R}^M$ the prior distribution over Ω, and by $U = [U_{km} = u(a_k, \omega_m)] \in \mathbb{R}^{K\times M}, V = [V_{km} = v(a_k, \omega_m)] \in \mathbb{R}^{K\times M}$ the sender's and the receiver's utilities, respectively. The sender's information structure is specified by a right stochastic matrix $\Pi = [\Pi_{mn} = \pi(s_n|\omega_m)] \in \mathbb{R}^{M\times N}$. The receiver's strategy is given by a right stochastic matrix $A = [A_{nk} = \alpha(a_k|s_n)] \in \mathbb{R}^{N\times K}$. Denote by $\mathbb{1}$ the all-one vector of a proper dimension depending on the context, and then $\Pi\mathbb{1} = \mathbb{1}, A\mathbb{1} = \mathbb{1}$.

In addition to the above, other helpful notations are as follows. e_i refers to the i-th elementary vector of a proper dimension depending on the context. For a vector w, $\text{diag}(w)$ denotes the diagonal matrix with w on its diagonal. For a square matrix W, $\text{diag}(W)$ denotes the vector containing its diagonal entries. For any two vectors $w, v \in \mathbb{R}^N$ of the same dimension, $w \succeq v$ (or $w \succ v$) indicates entry-wise relations: $w_i \geq v_i, \forall i \in [N]$. $\oslash$ denotes the Hadamard division (entry-wise): $w \oslash v = [w_i/v_i]_{i\in[N]}$. $\text{Tr}(W)$ denotes the trace of a square matrix W. W_j refers to the j-th column, and its transpose of W is denoted by W^T, while W' denotes its perturbation within the same domain specified by the context.

Define the prior matrix as $P = \text{diag}(p) \in \mathbb{R}^{M\times M}$. Given the players' strategies Π and A, the sender's expected payoff is $\sum_m p_m \sum_n \Pi_{mn} \sum_k A_{nk}U_{km} = \text{Tr}(P\Pi AU)$. Under the information structure Π, the receiver's posterior belief upon observing signal s_n is $\lambda_{mn} = \frac{p_m\Pi_{mn}}{\sum_{m'} p_{m'}\Pi_{m'n}}$. Define the belief system as $\Lambda = [\lambda_{mn}] \in \mathbb{R}^{M\times N}$. According to the Bayes rule shown above, the information structure and the belief system satisfies $\Lambda = P\Pi \oslash (\mathbb{1}\mathbb{1}^\mathsf{T}P\Pi)$. The receiver's strategy A is a best response to the posterior belief Λ, i.e., $\sum_m \lambda_{mn} \sum_k A_{nk}V_{km} \geq \sum_m \lambda_{mn} \sum_k A'_{nk}V_{km}$, for any $n \in [N]$, and any right stochastic matrix A'. Summing up all the equations above, we arrive at the following statement.

Proposition 1 (PBE in Matrix Form): For a finite communication game, a triple of matrices (Π, A, Λ) is a perfect Bayesian equilibrium if it satisfies

$$\text{Tr}(P\Pi AU) \geq \text{Tr}(P\Pi' AU), \forall \Pi' \in \mathbb{R}^{M\times N}_{\geq 0}, \Pi'\mathbb{1} = \mathbb{1}, \tag{6}$$

$$\text{diag}(AV\Lambda) \succeq \text{diag}(A'V\Lambda), \forall A' \in \mathbb{R}^{N\times K}_{\geq 0}, A'\mathbb{1} = \mathbb{1}, \tag{7}$$

$$\Lambda = P\Pi \oslash (\mathbb{1}\mathbb{1}^\mathsf{T}P\Pi), \tag{8}$$

$$\Pi\mathbb{1} = \mathbb{1}, A\mathbb{1} = \mathbb{1}, \Pi \in \mathbb{R}^{M\times N}_{\geq 0}, A \in \mathbb{R}^{N\times K}_{\geq 0}.$$

Proposition 1 clearly demonstrates that solving for PBE is challenging, as Hadamard division in the belief system makes the problem highly nonlinear. Fortunately, this nonlinearity created by Hadamard division can be bypassed using Bayesian plausibility for non-degenerate PBE, as shown later in Section IV. Finally, we conclude this section with the matrix representation of SPE in (4).

Proposition 2 (SPE in Matrix Form): For a finite communication game, a pair of matrices (Π, A) is a sender-preferred subgame perfect equilibrium if it satisfies

$$\begin{aligned} \max_{\Pi,A} \quad & \text{Tr}(P\Pi AU) \\ \text{s.t.} \quad & \text{diag}(AV\Lambda) \succeq \text{diag}(A'V\Lambda), \\ & \forall A' \in \mathbb{R}^{N\times K}_{\geq 0}, A'\mathbb{1} = \mathbb{1}, \\ & \Lambda^\mathsf{T} = P\Pi \oslash (\mathbb{1}\mathbb{1}^\mathsf{T}P\Pi), \\ & \Pi\mathbb{1} = \mathbb{1}, A\mathbb{1} = \mathbb{1}, \Pi \in \mathbb{R}^{M\times N}_{\geq 0}, A \in \mathbb{R}^{N\times K}_{\geq 0}. \end{aligned} \tag{9}$$

IV. Bayesian Plausibility and TSB Programming

This section develops the TSB programming approach to solve for PBE in Proposition 1, further enabling us to prove that PoT is tightly upper bounded by 1. We impose a standing assumption on the existence of non-degenerate PBE to secure the well-posedness of the proposed programming: for any finite communication games in this work, there exists at least one non-degenerate PBE.

Recall that Bayesian plausibility requires that $p = \sum_n \gamma_n \Lambda_n$, and $\gamma_n = \sum_m p_m \Pi_{mn}$ denotes the probability of generating s_n, which is a discrete counterpart to $\gamma(s)$ defined in Definition 1. For non-degenerate PBE, $\gamma_n > 0$ for all $n \in [N]$. Note that $\lambda_{mn} = \frac{p_m \Pi_{mn}}{\sum_{m'} p_{m'} \Pi_{m'n}}$, then $p_m \Pi_{mn} = \lambda_{mn} \gamma_n$. Hence, the sender's expected payoff can be rewritten using posterior beliefs Λ and $\Gamma := \mathrm{diag}(\gamma)$, as shown below:

$$\begin{aligned}\mathrm{Tr}(P\Pi AU) &= \sum_m p_m \sum_n \Pi_{mn} \sum_k A_{nk} U_{km} \\ &= \sum_n \sum_k \sum_m A_{nk} U_{km} p_m \Pi_{mn} \\ &= \sum_n \sum_k \sum_m A_{nk} U_{km} \lambda_{mn} \gamma_n = \mathrm{Tr}(AU\Lambda\Gamma).\end{aligned}$$

The above deduction actually gives an elementary proof of the one-to-one correspondence between information structure and the posterior distribution we discussed in (5). Meanwhile, as $\gamma \succ 0$, then $\mathrm{diag}(AV\Lambda) \succeq \mathrm{diag}(A'V\Lambda) \Leftrightarrow \mathrm{diag}(AV\Lambda\Gamma) \succeq \mathrm{diag}(A'V\Lambda\Gamma)$. Finally, one can see that both the sender's and the receiver's best response conditions involve the matrix product of Λ and Γ, creating another matrix representation of PBE presented in Theorem 1.

Define $Z = \Lambda\Gamma \in \mathbb{R}^{M\times N}$, then according to Bayesian plausibility, $Z\mathbb{1} = \Lambda\Gamma\mathbb{1} = \Lambda\gamma = p$. Another constraint on Z arises from the left stochasticity of Λ, i.e., $\mathbb{1}^\mathsf{T}\Lambda = \mathbb{1}^\mathsf{T}$, implying that $\mathbb{1}^\mathsf{T} Z = 1^\mathsf{T}\Lambda\Gamma = 1^\mathsf{T}\Gamma = \gamma^\mathsf{T} \succ 0$. Hence, $0 \prec Z^\mathsf{T}\mathbb{1} \prec \mathbb{1}$. Summarizing these constraints, we denote by $\mathcal{Z} := \{Z | Z \in \mathbb{R}^{M\times N}_{\geq 0}, Z\mathbb{1} = p, 0 \prec Z^\mathsf{T}\mathbb{1} \prec \mathbb{1}\}$ the set of Bayesian plausible matrices. We then arrive at the following theorem where PBE is characterized using the Z matrix.

Theorem 1: For a finite game, a pair matrices of (Z, A) is a non-degenerate perfect Bayesian equilibrium if it satisfies

$$\begin{aligned}&\mathrm{Tr}(AUZ) \geq \mathrm{Tr}(AUZ'), \forall Z' \in \mathcal{Z}, \\ &\mathrm{diag}(AVZ) \succeq \mathrm{diag}(A'VZ), \forall A' \in \mathbb{R}^{N\times K}_{\geq 0}, A'\mathbb{1} = \mathbb{1}, \\ &A^\mathsf{T}\mathbb{1} = \mathbb{1}, Z \in \mathcal{Z}. \end{aligned} \tag{10}$$

The significance of Theorem 1 is self-evident: no Hadamad division is involved, and (10) is a constrained bilinear programming with respect to Z and A. Intuitively, Z matrix transfers the equilibrium problem into the posterior belief space, eliminating the nonlinearity introduced by the Bayes rule [see (8)]. Due to the page limit, all proofs are deferred to the arXiv version.

Belief-Dominant Equilibrium: Even though Theorem 1 seems to be the light at the end of the tunnel, directly solving PBE using (10) is still daunting. It is cumbersome to vectorize Z and A and then transform (10) into a standard bilinear form. Since the vectorization concatenates row or column vectors of Z and A, the resulting vectors are no longer stochastic vectors, rendering many techniques in bilinear programming [15, Chapter 3.4] inapplicable. The following presents an equivalence between (10) and a two-stage-bilinear programming, where the optimal solutions to the first stage problem constitute the feasible set of the second stage programming.

Recall that the first inequality in (10) gives $\sum_{i\in[N]} a_i^\mathsf{T} U z_i = \mathrm{Tr}(AUZ) \geq \mathrm{Tr}(AUZ') = \sum_{i\in[N]} a_i^\mathsf{T} U z_i'$, where $a_i (a_i')$ and $z_i (z_i')$ are the i-th row vector of $A(A')$ and $i-$th column vector of $Z(Z')$, respectively. The trace inequality in (10) implies that the sender does not deviate from the equilibrium belief system Λ and the distribution Γ, as the resulting average payoff over every signal is optimal. Note that $z_i = \gamma_i \lambda_i$, λ_i is the i-th column of Λ, and consider the following constraints: $a_i^\mathsf{T} U \lambda_i \geq a_i^\mathsf{T} U \lambda_i'$, for any $i \in [N]$, $\lambda_i' \in \Delta([N])$. Compared to the trace, the newly introduced ones require the equilibrium belief itself to be optimal for each signal, fixing the receiver's move. The latter is stronger than the former. We refer to the PBE characterized by the stronger constraints as belief-dominant PBE, as the belief system Λ best responds to A and dominates all other beliefs.

Definition 2 (Belief-Dominant PBE): Non-degenerate PBE $(Z = \Lambda\Gamma, A)$ is said to be belief-dominant, if the belief system Λ best responds to A for each signal: $a_i^\mathsf{T} U \lambda_i \geq a_i^\mathsf{T} U \lambda_i'$, $a_i = (A^\mathsf{T})_i$, $\forall i \in [N]$, $\forall \lambda_i' \in \Delta([N])$.

The notion of belief dominance only applies to non-degenerate PBE, as the belief vector λ_i can be arbitrary when $\gamma_i = 0$ [see (3)] in degenerate cases. The following presents an example of belief-dominant PBE.

Example 1 (Non-degenerate and Belief-Dominant PBE): Consider a two-state, two-action, and two-signal case: $M = N = K = 2$. The sender's and the receiver's utility matrices are $U = [1, 0; 0, 1/2]$, $V = [1, 0; 0, 2]$. The prior is $p = (1/2, 1/2)$. Both parties prefer a_2 in state ω_2. The sender prefers a_1 in state ω_1, while the receiver is indifferent between two actions.

As shown in the utility matrices, the interests of the sender and the receiver are aligned. Hence, one special perfect Bayesian equilibrium strategy for the sender is the so-called truth-telling strategy: $\Pi = I$, also referred to as the separating equilibrium [6, Chapter 8]. As a non-degenerate PBE, the separating equilibrium is given by

$$\Pi = \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Gamma = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}, A = \begin{bmatrix} 1-\epsilon & \epsilon \\ 0 & 1 \end{bmatrix},$$

where $\epsilon \in [0, 1]$. The ϵ entries in A are due to the fact that the receiver is indifferent between two actions when it observes s_1 and realizes that the state is ω_1. In other words, there is a continuum of non-degenerate PBE. Direct calculation gives that those PBE with $\epsilon \in [0, 2/3]$ is belief-dominant.

Two-Stage-Bilinear Programming For belief-dominant

PBE, the constraints in (10) reduces to

$$\begin{aligned} &a_i^\mathsf{T} U\lambda_i \geq a_i^\mathsf{T} U\lambda_i', \lambda_i^\mathsf{T}\mathbb{1} = \lambda_i'^\mathsf{T}\mathbb{1} = 1, \lambda_i, \lambda_i' \geq 0, \\ &a_i^\mathsf{T} V\lambda_i \geq a_i'^\mathsf{T} V\lambda_i, a_i^\mathsf{T}\mathbb{1} = a_i'^\mathsf{T}\mathbb{1} = 1, a_i, a_i' \geq 0, \end{aligned} \tag{11}$$

for any $i \in [N]$. The benefit of considering these constraints is that all decision variables involved are from the probability simplex, leading to the helpful results below.

Lemma 1: A pair $(\{a_i\}, \{\lambda_i\})$ constitutes a feasible point to (11) if and only if there exists a pair $(\{x_i\}, \{y_i\})$ such that $(\{a_i\}, \{\lambda_i\}, \{x_i\}, \{y_i\})$ solves the bilinear programming

$$\begin{aligned} \max_{a_i,\lambda_i,x_i,y_i} &\sum_{i\in[N]} a_i^\mathsf{T} U\lambda_i + \sum_{i\in[N]} a_i^\mathsf{T} V\lambda_i - \sum_{i\in[N]} x_i - \sum_{i\in[N]} y_i \\ \text{s.t. } &U^\mathsf{T} a_i \preceq x_i\mathbb{1}, V\lambda_i \preceq y_i\mathbb{1}, \\ &a_i \geq 0, \lambda_i \geq 0, a_i^\mathsf{T}\mathbb{1} = 1, \lambda_i^\mathsf{T}\mathbb{1} = 1. \end{aligned} \tag{12}$$

Corollary 1: For any solution $(\{a_i\}, \{\lambda_i\}, \{x_i\}, \{y_i\})$ to the bilinear programming (12), if there exists a $\gamma \in \Delta([M])$ such that $\Lambda\gamma = p$, $\Lambda_i = \lambda_i$, then (Λ, γ) is Bayesian plausible. Under the corresponding information structure, the sender's (receiver's) expected payoff under signal s_i is x_i (y_i).

The proof of Corollary 1 rests on the fact that the solution quadruple to (12) satisfies the equations: $a_i^\mathsf{T} U\lambda_i = x_i$, $a_i^\mathsf{T} V\lambda_i = y_i$. As one can see from Corollary 1, an optimal posterior distribution γ should maximize the expected payoff of all signals: $\sum_i \gamma_i x_i$. This observation leads to Theorem 2, where solutions to the bilinear programming in (12) constitute a feasible set to another bilinear programming.

Theorem 2 (Two-Stage-Bilinear Programming): For any solution $(Z = (\Lambda, \gamma), A)$ to (10) that is belief-dominant, it is also a solution to (13). Conversely, if $\{(A^\mathsf{T})_i, \Lambda_i, x_i, y_i, \gamma_i\}$ solves (13) and satisfies $\Lambda\gamma = p$, $\gamma^\mathsf{T}\mathbb{1} = 1$, $\gamma \succ 0$, then (Λ, γ, A) solves (10).

$$\max_{\gamma_i, x_i} \sum_i \gamma_i x_i, \text{ s.t. } \{(A^\mathsf{T})_i, \Lambda_i, x_i, y_i\} \text{ solves (12).} \tag{13}$$

Before inspecting the tightness of the upper and lower bounds, we introduce a finite-time algorithm to find the exact solution to (13). The first step is to identify the feasible set characterized by (12). One can see from the proof and (11) that the purpose of bilinear programming (12) is to enumerate all solutions of the bimatrix game (U, V). Prior works [16], [17] have established finite-time algorithms (with exponential complexity) to enumerate the exact solutions. An online solver, named `lrs` (lexicographic reverse search), is offered by [17]. Consider the binary communication game in Example 1. `lrs` returns three solution tuples represented as (a, λ, x): $\{(1/2, 1/2), (1/3, 2/3), 1/3\}$, $\{(0, 1), (0, 1), 1/2\}$, and $\{(1, 0), (1, 0), 1\}$. We now turn to the bilinear programming in (13). Since $N = 2$, we only need to keep two solution tuples as the feasible points. As $1/3 < 1/2 < 1$, it is natural to drop the solution $\{(1/2, 1/2), (1/3, 2/3), 1/3\}$ and keep the other two. In this case, the belief matrix becomes $\Lambda = I$, and hence, $\gamma = p = (1/2, 1/2)$. Since the feasible set for the variable γ is a singleton, then the optimal solution is exactly the truth-telling strategy in Example 1. Note that `lrs` only returns extreme equilibria [17], which explains why Example 1 presents a continuum of PBE, whereas `lrs` only gives the above three. Yet, this technical nuance does not affect the objective value in (13). A more detailed discussion is presented in the arXiv version.

The Tight Upper Bound To evaluate the PoT, we first transfer the SPE in (9) into the posterior belief space, which is given by the following programming:

$$\begin{aligned} \max_{\{\gamma_i, \lambda_i, a_i\}} &\sum_i a_i^\mathsf{T} U\lambda_i\gamma_i \\ \text{s.t. } &a_i^\mathsf{T} V\lambda_i \geq a_i'^\mathsf{T} V\lambda_i, a_i^\mathsf{T}\mathbb{1} = a_i'^\mathsf{T}\mathbb{1} = 1, a_i, a_i' \geq 0, \\ &\sum_i \lambda_i\gamma_i = p. \end{aligned} \tag{14}$$

Comparing (14) and (11), one can see that PBE admits one more constraint regarding λ. Hence, PoT ≤ 1. The following introduces a special class of communication games, referred to as strictly Bayesian-posterior competitive games, for which we prove that the upper bound is attained.

Definition 3 (Strictly Bayesian-posterior competitiveness): A game with payoffs (U, V) is strictly Bayesian-posterior competitive if for all $a, a' \in \Delta(A)$, $\lambda, \lambda' \in \Delta(\Omega)$, $a^\mathsf{T} U\lambda - a'^\mathsf{T} U\lambda'$ and $a'^\mathsf{T} V\lambda' - a^\mathsf{T} V\lambda$ have the same sign.

Theorem 3 (Tightness of the Upper Bound): Assuming that belief-dominant equilibrium exists for some strictly Bayesian-posterior competitive game, then PoT $= 1$.

V. Quadratic Games and Tight Lower Bound

Note that (10) and (13) give a TSB characterization of PBE, which is instrumental in showing the tightness of the upper bound in finite games. However, the bilinear programming does not reveal the tightness of the lower bound. This section presents a case study on the PoT in a particular continuous game: quadratic communication game (QCG), where PoT can be arbitrarily close to zero.

QCG consists of continuous state, signal, and action spaces: $\Omega = \mathcal{S} = \mathcal{A} = [0, 1]$, as well as quadratic utilities: $u(a, \omega) = -(a - \omega - b)^2, v(a, \omega) = -(a - \omega)^2$. The bias term $b > 0$ denotes the misalignment between two players' interests: as $b \to 0$, the two are more aligned. The receiver tries to guess where the actual state is (minimizing the error) based on the signal from the sender, who tries to mislead the receiver to somewhere else (specified by the offset b). The prior is the uniform distribution denoted by $p = \text{unif}(0, 1)$.

PBE in Signaling An important finding in [8] is that all PBE in QCG are partition equilibria. Given a constant $b > 0$, there exists a positive integer $N(b) = \lfloor -\frac{1}{2} + \frac{1}{2}(1 + \frac{2}{b})^{1/2} \rfloor$ ($\lfloor \cdot \rfloor$ is the ceiling function) such that there exists a PBE for every $N \in [N(b)]$. The equilibrium information structure is in the form of partition signaling: for any $N \in [N(b)]$, there exists a sequence $0 = k_0 < k_1 < \ldots < k_N = 1$ such that

$$\pi(\cdot|\omega) = \text{unif}(k_i, k_{i+1}), \text{if } \omega \in (k_i, k_{i+1}). \tag{15}$$

In the partition equilibria, the sender randomly samples a signal from the sub-interval within which the true state falls, telling a half-truth to the receiver. One can clearly see from (15) that the more nearly players' interests coincide (the closer b approaches zero), the finer partition there can be (the larger $N(b)$). On the contrary, as $b \to \infty$, $N(b)$ eventually

falls to unity, and the sender would transmit uninformative signals to the receiver. Direct calculation shows that the watershed is $1/4$: as b exceeds $1/4$, $\pi(\cdot|\omega) = p$ for all ω. For the rest of this section, we assume $b \in (0, 1/4)$.

We now turn to the sender's PBE payoff, i.e., U^{CS}. [8, Theorem 1] states that under the information structure (15), $U^{CS} = -\sum_{i\in[N]} \mathrm{Var}_{\mathrm{unif}(k_{i-1},k_i)} - b^2$, where $\mathrm{Var}_{\mathrm{unif}(k_i,k_{i+1})}$ denotes the variance of the uniform distribution over $[k_{i-1}, k_i]$. The equilibrium partition of number $N \in [N(b)]$, as shown in [8], is $k_i = i/N + 2bi(i-N)$, $i \in [N]$. Hence, a direct calculation gives

$$U^{CS} = -\frac{1}{12N^2} - \frac{b^2(N^2-1)}{3} - b^2. \tag{16}$$

SPE in Persuasion The calculation of SPE in QCG rests on the backward induction in (5). Given a posterior belief λ, the best response is $\hat{a}(\lambda) = \arg\max_a \mathbb{E}_{\omega\sim\lambda}[-(a-\omega)^2] = \mathbb{E}_\lambda[\omega]$. The sender's expected payoff under λ is $\hat{u}(\lambda) = \mathbb{E}_{\omega\sim\lambda} u(\hat{a}(\lambda), \omega) = -\mathrm{Var}_\lambda - b^2$, where Var_λ denotes the variance of the random variable $\omega \sim \lambda$. Finally, U^{OP} is the optimal value of the following problem:

$$\max_{\tau\in\Delta(\Delta([0,1]))} \mathbb{E}_\tau[-\mathrm{Var}_\lambda - b^2], \text{ s.t. } \int \lambda\tau(d\lambda) = p.$$

If we choose λ as a Dirac measure $\delta(\omega)$, for $\omega \in [0,1]$, then the variance term is zero. Hence, the optimal value is $U^{OP} = -b^2$. The interpretation is that the sender opts for a truth-telling strategy, i.e., $s = \omega$, even though the incentive bias exists $b > 0$. Consequently, the receiver's belief collapses to its true state $\lambda(\cdot|\omega) = \delta(\omega)$. Mathematically, this truth-telling equilibrium is due to the fact that $\hat{u}(\lambda)$ is convex in λ, and the reader is referred to [7, Section V] for more details, where authors consider a lobby game similar to our setting.

The Tight Lower Bound With all the results above, we now address the lower bound of PoT. Note that for the simplicity of exposition, we construct non-positive utilities in QCG, violating the non-negativity assumption in Section II. If blindly computing $\frac{U^{CS}}{U^{OP}}$, one would arrive at the opposite conclusion. Therefore, we prove that PoT converges to zero by showing that U^{OP} converges to zero (the maximum) at a higher order than U^{CS} does, as $b \to 0$. This higher-order convergence indicates that OP significantly outperforms CS.

Theorem 4 (Tightness of the Lower Bound): Consider the quadratic communication game of the incentive bias $b > 0$, PoT converges to 0, as b tends to 0.

Remark 1 (Half-Truth still Hurts.): The opaque information disclosure, compared to the transparent, creates "friction" in information transmission. As one can see from the partition equilibria in (15), the informativeness of the signaling is reflected by the width of each sub-interval $d_i = k_{i+1} - k_i$. The finer the partition is, the smaller d_i is, and the more confident the receiver is about the true state. As $b \to 0$, and d_i shrinks, the half-truth gets closer to the truth. Yet, the half-truth still hurts: the signal bears randomness (unlike the deterministic signal in OP), even though the two players' interests coincide. The resulting U^{CS} exhibits a first-order convergence.

VI. Conclusion

This work has introduced the notion of *price of transparency* (PoT) to quantify the cost or benefit of information disclosure in strategic interactions. It allows for the assessment of the sender's tradeoffs when adhering to ethical standards that require transparency in information disclosure. We have observed that counterintuitively, choosing transparency can yield a payoff no less than that under an opaque information structure, with PoT values ranging between 0 and 1. We have developed a two-stage-bilinear programming approach (10) using Bayesian plausibility to solve for the perfect Bayesian equilibrium. Furthermore, this programming approach has enabled us to show the upper bound is attainable for strictly Bayesian-posterior competitive games. Additionally, we have constructed quadratic games where PoT can be arbitrarily close to 0. The tight lower bound implies that the sender can be plagued by the lack of transparency. One future direction is to investigate a class of incentive mechanism design problems where the designer creates incentives for agents to adhere to transparency.

References

[1] J. Pawlick, E. Colbert, and Q. Zhu, "A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy," *ACM Comput. Surv.*, vol. 52, aug 2019.

[2] Y. Pan, T. Li, H. Li, T. Xu, Z. Zheng, and Q. Zhu, "A first order meta stackelberg method for robust federated learning," *arXiv preprint arXiv:2306.13800*, 2023.

[3] Y. Ge, T. Li, and Q. Zhu, "Scenario-agnostic zero-trust defense with explainable threshold policy: A meta-learning approach," in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, 2023.

[4] T. Li, G. Peng, Q. Zhu, and T. Başar, "The Confluence of Networks, Games, and Learning a Game-Theoretic Framework for Multiagent Decision Making Over Networks," *IEEE Control Systems*, vol. 42, no. 4, pp. 35–67, 2022.

[5] T. Li, Y. Zhao, and Q. Zhu, "The role of information structures in game-theoretic multi-agent learning," *Annual Reviews in Control*, vol. 53, pp. 296–314, 2022.

[6] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1991.

[7] E. Kamenica and M. Gentzkow, "Bayesian Persuasion," *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011.

[8] V. P. Crawford and J. Sobel, "Strategic Information Transmission," *Econometrica*, vol. 50, no. 6, p. 1431, 1982.

[9] O. Gossner, "Comparison of Information Structures," *Games and Economic Behavior*, vol. 30, no. 1, pp. 44–63, 2000.

[10] J. R. Green and N. L. Stokey, "A two-person game of information transmission," *Journal of Economic Theory*, vol. 135, no. 1, pp. 90–104, 2007.

[11] A. Rubinstein, "Honest signaling in zero-sum games is hard, and lying is even harder," *arXiv*, 2015.

[12] U. Bhaskar, Y. Cheng, Y. K. Ko, and C. Swamy, "Hardness results for signaling in bayesian zero-sum and network routing games," in *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, (New York, NY, USA), p. 479–496, Association for Computing Machinery, 2016.

[13] S. Dughmi, "On the hardness of designing public signals," *Games and Economic Behavior*, vol. 118, pp. 609–625, 2019.

[14] T. Li and Q. Zhu, "Commitment with Signaling under Double-sided Information Asymmetry," *arXiv preprint: 2212.11446*, 2022.

[15] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory, 2nd Edition*. Society for Industrial and Applied Mathematics, 1998.

[16] G. Gallo and A. Ülkücü, "Bilinear programming: An exact algorithm," *Mathematical Programming*, vol. 12, no. 1, pp. 173–194, 1977.

[17] D. Avis, G. D. Rosenberg, R. Savani, and B. v. Stengel, "Enumeration of Nash equilibria for two-player games," *Economic Theory*, vol. 42, no. 1, pp. 9–37, 2010.