Strategic underdisclosure and the value of jammed signals in evidence games[*]

PRELIMINARY

Orestis Vravosinos[†]

April 8, 2024

Abstract

Keywords: information transmission, disclosure, persuasion
JEL classification codes: D82, D83

[*] I am grateful to Agata Farina, Alistair Barton, Erik Madsen for his continued guidance and insightful comments. This work has also significantly benefited from discussions with Debraj Ray. I also thank Agata Farina and participants in the NRET seminar at NYU for helpful comments and discussions.

[†] New York University; e-mail: orestis.vravosinos@nyu.edu.

1 Introduction

2 The model

There are two players, a sender (she) and a receiver (he). The sender's bi-dimensional type (e,t) has with full support density $f : [0,1]^2 \to \mathbb{R}_+$. The sender is privately informed of dimension e. t is not observed by either player. e is the agent's *evidence*. That is, an agent of type (e,t) can prove to the principal that her e is at least r for any $r \in [0,e]$ by presenting evidence $r \in [0,e]$. If she reveals $r < e$, we say that she partially reveals her evidence. However, for no $r \in [0,1)$ can she prove that her e is not higher than r; in other words, she cannot prove that she is not withholding evidence. The second dimension of the sender's type t is her *talent*. The game proceeds as follows:

1st stage: The receiver chooses whether to test the sender by paying a cost $c \geq 0$. If he tests her, then he receives a (deterministic) mixed signal $\sigma(e,t) \in [0,1]$ of the sender's type, where $\sigma : [0,1]^2 \to [0,1]$ is continuous and increasing in both arguments. Denote by $\tau \in \{0,1\}$ his choice, where $\tau = 1$ (resp. $\tau = 0$) means that he tests (resp. does not test) the sender. Also, for every possible test result $s \in [0,1]$, define $\widehat{t}_s(e)$ implicitly given by $\sigma(e,\widehat{t}_s(e)) = s$.

2nd stage: After observing the receiver's decision to test or not (but not the test result), the sender chooses how much evidence $r \in [0,e]$ to present to the receiver.[1]

3rd stage: The receiver sees the amount of evidence r presented by the sender and, if he chose to perform the test, the test result $s = \sigma(e,t)$ and then takes an action $a \in [0,1]$.

Payoffs: The sender's payoff is equal to a. The receiver's payoff is given by $u_{e,t}(a) - \tau c$. $u_{e,t}(a)$, his gross payoff, is single-peaked in a with $\arg\max_{a\in[0,1]} u_{e,t}(a) = v(e,t)$, where $v : [0,1]^2 \to [0,1]$ is non-decreasing and continuous in both arguments. For example, $u_{e,t}(a) := -(v(e,t) - a)^2$, where $v(e,t)$ is non-decreasing and continuous in both e and t.

A strategy profile (τ,r_0,r_1,a_0,a_1) is comprised of the receiver's testing decision $\tau \in \{0,1\}$, the sender's evidence provision strategy $e \to r_1(e)$ (resp. $e \to r_0(e)$) if the receiver chooses to test (resp. not to test), and the receiver's action strategy $r \to a_0(r)$ (resp. $(r,s) \to a_1(r,s)$) if the receiver chooses to test (resp. not to test).

Equilibrium with lying costs. The sender's payoff is equal to $a - k(e - r)^2$. The receiver's payoff is given by $-\{a - [w_u t + (1 - w_u)e]\}^2 - \tau c$. $\sigma(e,t) = w_\sigma t + (1 - w_\sigma)e$.

If the sender does test and $w_\sigma < w_u$, then there are incentives to hide evidence.

In equilibrium, $a_1(r,s) = \mathbb{E}[w_u t + (1 - w_u)e|r,s]$. Take $e_1 < e_2$ and assume that

[1] In the sender's problem of choosing r, only the decision of the receiver to test or not matters, and not whether the receiver observes the signal (if the sender tests) in the 2nd or 3rd stage. Also, in case the receiver mixes in his decision to test or not, the sender observes the outcome of the mixing before choosing r. Thus, there is no need to consider mixing in the testing decision.

$r_1(e_1) = r_1(e_2) \leq e_1$. Then, e_1's payoff is equal to

$$\int_s \mathbb{E}[w_u t + (1 - w_u)e | r_1(e_2), s] dF_{e_1}(s) - k(e_1 - r_1(e_1))^2,$$

while e_2's payoff is equal to

$$\int_s \mathbb{E}[w_u t + (1 - w_u)e | r_1(e_2), s] dF_{e_2}(s) - k(e_2 - r_1(e_1))^2.$$

Definition 1. A strategy profile (τ,r_0,r_1,a_0,a_1), where $\tau \in \{0,1\}$, $r_0,r_1 : [0,1]^2 \to [0,1]$, $a_0 : [0,1] \to [0,1]$, and $a_1 : [0,1]^2 \to [0,1]$ is an equilibrium if

(i) given receiver beliefs $\mu_0(r)$ (resp. $\mu_1(r,s)$) when he does not test, a_0 (resp. a_1) maximizes the receiver's payoff,

(ii) r_0 (resp. r_1) maximizes the sender's payoff given a_0 (resp. a_1) when the receiver does not test (resp. does test),

(iii) τ maximizes the receiver's payoff given (r_0,r_1,a_0,a_1),

(iv) μ_0 and μ_1 respect Bayes rule, and

(v) $\mu_0(r)$ assigns probability 1 to $e = r$ and density $f(t|r)$ to t for $r \notin r_0([0,1]^2)$,

(vi) $\mu_1(r,s)$ assigns probability 1 to $(e,t) = (r,\widehat{t}_s(r))$ for $r \notin r_1([0,1]^2)$ and s such that $\sigma(e,t) = s$.

Definition 2. Define $u_{e,f(t|e)}(a) := \int_0^1 u_{e,t}(a) f(t|e)dt$, where $f(t|e) := f(e,t)/\int_0^1 f(e',t)de'$ the distribution of t conditional on e. We say that:

(i) Evidence is a priori good if $\arg\max_{a\in[0,1]} u_{e,f(t|e)}(a)$ is increasing in e.

(ii) Evidence is a priori bad if $\arg\max_{a\in[0,1]} u_{e,f(t|e)}(a)$ is decreasing in e.

(iii) Evidence is post-testing good if $\arg\max_{a\in[0,1]} u_{e,\widehat{t}_s(e)}(a)$ is increasing in e for every test result $s \in [0,1]$.

(iv) Evidence is post-testing bad if $\arg\max_{a\in[0,1]} u_{e,\widehat{t}_s(e)}(a)$ is decreasing in e for every test result $s \in [0,1]$.

Remark: $u_{e,\widehat{t}_s(e)}(a)$ is the receiver's gross payoff from choosing action a when the sender's type is $(e,\widehat{t}_s(e))$.

Lemma 1. If the receiver tests the sender in the 1st stage, then

(i) if evidence is post-testing good, then there exists an essentially unique equilibrium in the post-testing subgame: every sender (e,t) reveals all her evidence in the 2nd stage, $r^*(e,t) = e$, and the receiver's strategy in the 3rd stage is $a^*(r,s) = v(r,\widehat{t}_s(r))$, while

(ii) if evidence is post-testing bad, then there exists a unique equilibrium in the post-testing subgame: every sender (e,t) reveals no evidence in the 2nd stage, $r^*(e,t) = 0$, and the receiver's on-path strategy in the 3rd stage specifies $a^*(0,s) = \arg\max_{a\in[0,1]} u_{f(e,t|s)}(a)$.

Proof of Lemma 1 Fix any $r,s \in [0,1]^2$ such that $\sigma(r,0) \leq s \leq \sigma(r,1)$ and observe that in any equilibrium, after seeing evidence r and test score s, the the receiver must hold beliefs with $\operatorname{supp} \mu_1(r,s) \subseteq \{(e,t) \in [0,1]^2 : \sigma(e,t) = s \wedge e \geq r\}$.

(i) Then, given that evidence is post-testing good and the receiver's payoff is single-peaked in a, in equilibrium the receiver's strategy must satisfy

$$a_1(r,s) \geq \arg\max_{a \in [0,1]} u_{r,\widehat{t}_s(r)}(a). \tag{1}$$

Also, in any equilibrium where multiple sender types with test score s present the same amount of evidence, there exists (at least) one among them (e^*,t^*) that earns payoff lower than $\arg\max_{a\in[0,1]} u_{e^*,t^*}(a)$. Particularly, if $P \coloneqq \{(e,t) \in [0,1]^2 : \sigma(e,t) = s \wedge r_1(e,t) = s\}$ has cardinality $|P| > 1$, then $a_1(r_1(e^*,t^*),s) < \arg\max_{a\in[0,1]} u_{e^*,t^*}(a)$ for $(e^*,t^*) \in P$ with e^* close enough to $\sup_{(e,t)\in P} e$. But this combined with (1) gives

$$a_1(r_1(e^*,t^*),s) < \arg\max_{a \in [0,1]} u_{e^*,t^*}(a) \leq a_1(e^*,s),$$

so (e^*,t^*) has a profitable deviation to presenting evidence e^* instead. Therefore, there is no equilibrium where multiple sender types with the same test score present the same amount of evidence, and the result follows.

(ii) Given that evidence is post-testing bad and the receiver's payoff is single-peaked in a, in equilibrium the receiver's strategy must satisfy

$$a_1(r,s) \leq \arg\max_{a \in [0,1]} u_{r,\widehat{t}_s(r)}(a). \tag{2}$$

In any equilibrium where two distinct sender types $(\underline{e},\overline{t})$ and $(\overline{e},\underline{t})$, $\overline{e} > \underline{e}$, with test score $\sigma(\underline{e},\overline{t}) = \sigma(\overline{e},\underline{t}) = s$ earn different payoffs it must be that $(\overline{e},\underline{t})$ earns a higher payoff than $(\underline{e},\overline{t})$. Namely, by $(\overline{e},\underline{t})$'s best-responding and since $r_1(\underline{e},\overline{t}) \leq \underline{e} < \overline{e}$ it follows that $a_1(r_1(\overline{e},\underline{t}),s) \geq a_1(r_1(\underline{e},\overline{t}),s)$, and thus (given that $a_1(r_1(\overline{e},\underline{t}),s) \neq a_1(r_1(\underline{e},\overline{t}),s)$),

$$a_1(r_1(\overline{e},\underline{t}),s) > a_1(r_1(\underline{e},\overline{t}),s). \tag{3}$$

Therefore, sender optimality implies that $\{(e,t) : \sigma(e,t) = s \wedge r_1(e,t) = r_1(\underline{e},\overline{t})\} \subseteq \{(e,t) : \sigma(e,t) = s \wedge e \in [\underline{e},\overline{e})\}$. Thus, given that evidence is post-testing bad and the receiver's payoff is single-peaked in a, receiver optimality implies that

$$a_1(r_1(\underline{e},\overline{t}),s) \leq \arg\max_{a \in [0,1]} u_{\underline{e},\overline{t}}(a)),$$

which combined with (3) implies . Thus, in any equilibrium, all sender types with the same test score earn the same payoff.