CENTER FOR ADVANCED RESEARCH IN FINANCE

CARF Working Paper

CARF-F-496

Unique Information Elicitation

Hitoshi Matsushima
University of Tokyo

Shunya Noda
University of British Columbia

October 26, 2020

CARF is presently supported by The Dai-ichi Life Insurance Company, Limited, Nomura Holdings, Inc., Sumitomo Mitsui Banking Corporation, Mizuho Financial Group, Inc., MUFG Bank, Ltd., The Norinchukin Bank and The University of Tokyo Edge Capital Partners Co., Ltd. This financial support enables us to issue CARF Working Papers.

Unique Information Elicitation[1]

Hitoshi Matsushima[2]

University of Tokyo

Shunya Noda[3]

University of British Columbia

October 26, 2020

Abstract

This study investigates the unique information elicitation problem. A central planner attempts to elicit correct information from multiple informed agents through mutual monitoring. There is a severe restriction on incentive devices: we assume neither public monitoring technology nor allocation rule is available; thus, the central planner only uses monetary payment rules. It is well-known that if all agents are selfish, it is impossible to elicit information as a unique equilibrium. We consider an epistemological possibility that some agents could be motivated by an intrinsic preference for honesty, while we allow that honest agents are mostly motivated by monetary interest. We prove that, once we introduce an epistemic type space that allows agents to be (weakly) honest, then the impossibility theorem reduces to a knife-edge case: The central planner can elicit correct information from agents as a unique Bayes Nash equilibrium behavior if and only if it is never common knowledge that all agents are selfish.

Keywords: Unique Information Elicitation, Restricted Incentive Devices, Mutual Monitoring, Preference for Honesty, Common Knowledge of All Agents' Selfishness.
JEL Classification Numbers: C72, D71, D78, H41

---

[1] This study is a drastic extension of the theory part of Matsushima and Noda (2020), where we added an epistemological framework for considering common knowledge. This study was supported by a grant-in-aid for scientific research (KAKENHI 20H00070) from the Japan Society for the Promotion of Science (JSPS), the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese government, and the Social Sciences and Humanities Research Council of Canada..

[2] Department of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: hitoshi [at] e.u-tokyo.ac.jp

[3] Vancouver School of Economics, University of British Columbia, 6000 Iona Dr, Vancouver, BC V6T 1L4, Canada. E-mail: shunya.noda [at] gmail.com

1. Introduction

This study investigates a mechanism design problem in which a central planner attempts to elicit correct information from agents. The central planner needs to know which state of the world actually occurs, but she (or he) is not informed of it. However, there exists an agent who is fully informed of it. Hence, the central planner attempts to design a mechanism to incentivize this agent to announce the state truthfully. This study clarifies the possibility that the central planner successfully elicits correct information *even when there are severe restrictions on incentive devices available*.

Several previous works such as the principal-agent problem with hidden information [4] have assumed public monitoring technology in which, through the observation of an ex-post public signal, the central planner confirms at least in part whether an agent announces truthfully. On the contrary, this study assumes that there exists no such public monitoring technology available.

Other previous works in the literature such as auction and implementation theory[5] assumed that there exist multiple tools for incentives such as resource allocations with which the central planner extracts correct information by letting agents self-select from multiple menus of allocations. On the contrary, the central planner in this study cannot use any such allocation device besides monetary transfers: *she is only permitted to use a message-contingent payment rule*.

To overcome the difficulty due to these restrictions, the central planner will listen to the messages from multiple agents who have the same information as each other and having them *mutually monitor* with each other. However, for such mutual monitoring to function, the central planner still needs to overcome another challenge in incentives, i.e., the multiplicity of equilibria due to the coordination failure. Hence, this study clarifies the possibility of *unique information elicitation*, implying that the central planner elicits correct information through agents' unique equilibrium behavior.

The mechanism design literature has traditionally assumed that it is common knowledge that all agents are selfish and only concerned about their respective monetary

---

[4] See Salanié (1997) for surveys on principal-agent problem with hidden information.
[5] See Krishna (2009) for a survey on auction theory and Maskin and Sjöström (2002) for surveys on implementation theory.

and material interests. With this assumption, it is inevitable to have severe multiplicity of equilibria in our problem, because agents' monetary interest is independent of the state: the set of all equilibria is the same across states in such a standard model-setting. However, this common knowledge of all agents' selfishness is clearly an unrealistic assumption: real people are often likely to have non-selfish motives such as ethical concerns, depending on the context of the actual situation that the abstract model describes. Hence, the statement derived from this assumption is useful only if it is robust against contamination of non-selfish motives. We prove that this is not the case at all as for the unique information elicitation problem: Once we incorporate a psychological motive and an epistemological type space, we obtain a drastically different result from a standard setting where non-selfish agents are excluded completely.

We consider the possibility that an agent is not selfish and instead motivated by *intrinsic preference for honesty* as well as monetary interest. We however permit that a majority of agents are selfish: we just eliminate the common knowledge assumption on all agents' selfishness. This study then demonstrates a surprising result: *the central planner can overcome this multiplicity and elicit correct information from agents through unique Bayes Nash equilibrium (BNE) behavior, if and only if it never happens to be common knowledge that all agents are selfish.* This is a very powerful and profound statement: it provides a theoretical basis that a person who commits wrongdoing in the world can be caught only by testimony, ex post facto (limited incentive tools), or without any means of proof (no provability).

In reality, many empirical and experimental studies have indicated that human beings are not purely motivated with monetary payoffs but real-world people have intrinsic preferences for honesty. Abeler et al. (2019) provided a detailed meta-analysis: they combined data from 90 studies involving more than 44,000 subjects across 47 countries and showed that subjects forwent a large fraction of potential gain from lying. In addition, various papers in behavioral economics and decision theory have modeled preferences for honesty, such as a cost of lying (e.g., Ellingsen and Johannesson, 2004; Kartik, 2009), a reputational cost (e.g., Mazar, Amir, and Ariely, 2008), and guilt aversion (e.g., Charness and Dufwenberg, 2006). Accordingly, it is rather unrealistic to assume that all participants of mechanisms are purely interested in their monetary payoffs.

This study however allows the case that there exists an honest agent only exceptionally. The influences of preferences for honesty on decision making can be arbitrarily small even if she is honest: We allow that honest agents are mostly motivated by monetary interest. We only rule out the case in which it happens to be common knowledge that there exists no honest agent. Just by eliminating the common knowledge of all agents' selfishness, we establish the possibility result.

Epistemologically, this study does not assume that agents expect a possibility that there exists an honest agent. This study even allows agents to have mutual knowledge that all agents are selfish. Despite these weaknesses in honesty, the central planner can incentivize all agents whether selfish or honest to announce truthfully as unique BNE behavior.

The design of the payment rule in this study has the following characteristics. First of all, each agent is required to announce not a state but a *distribution of the state*, even if she (or he) is fully informed of the state: she can fine-tune her announcement and payoff continuously. Second, each agent always prefers announcing the *same distribution* as what the other agents announce in expectation, whenever she is selfish. On the other hand, an honest agent is driven to be *more honest* than a selfish agent due to an intrinsic preference for honesty. Based on these characteristics, all agents come to expect a possibility that an agent is driven to be more honest, which drives all agents into a tail-chasing competition toward honest reporting.

We design the payment rule as a version of the *quadratic scoring rule (Brier, 1950)*, which describes the distance between agents' messages. The quadratic scoring rule is one of the standard methods of mechanism design in partial implementation with asymmetric information.[6] This study suggests that this method is a powerful solution not only for partial implementation but also for *unique implementation*.

The equilibrium analysis of the game under the presence of behavioral agents and incomplete information itself has a long history. For example, Kreps et al (1982) studied how the existence of behavioral agents changes the equilibria of finitely repeated games. Postlewaite and Vives (1987), Carlsson and van Damme (1993), and Morris and Shin

[6] See Cooke (1991) for a survey of scoring rules. For its applications to mechanism design, see Johnson et al (1990), Matsushima (1990; 1991; 1993; 2007), Aoyagi (1998), and Miller et al (2007), for example.

(1998) studied how incomplete information shrinks the set of equilibria. These previous studies focus on the analysis of given games. In contrast, our focus is on the design of mechanisms, which fully take advantage of the possibility of behavioral agents and incomplete information.

More recently, several studies have also investigated mechanism design under the presence of honest agents (Matsushima, 2008a; 2008b; Dutta and Sen, 2012; Kartik et al., 2014). The result of this study is novel in the following two senses. First, this paper studies an information elicitation problem in which the central planner makes no resource allocation. Consequently, the central planner cannot take advantage of the relationship between the realized state and agents' preferences, and therefore, the implementation must be purely based on preferences for honesty. Second, we do not assume the existence of honest agents—the only assumption we need is the event "all agents are selfish" is not common knowledge. We show that, even in such an environment, unique information elicitation is possible.

A number of studies have extended the scoring rule of Brier (1950) to a setting in which a central planner collects information from a group of agents (e.g., Dasgupta and Ghosh, 2013; Prelec, 2004; Miller et al., 2005; Kong and Schoenebeck, 2019). Previous studies have assumed that all agents are selfish, and therefore, have suffered from the multiplicity of equilibria. In contrast, we prove that the impossibility of unique information elicitation is a knife-edge result: whenever selfishness is not common knowledge, unique information elicitation is possible.

The organization of this study is as follows. Section 2 shows the model. Section 3 shows the main theorem. Section 4 demonstrates an example that outlines the logic behind this theorem. Section 5 shows the complete proof of the theorem. Section 6 shows applications and extensions. Section 7 concludes.

2. The Model

This study investigates a situation in which a central planner attempts to elicit information from multiple agents correctly. Let $N = \{1, 2, ..., n\}$ denote the finite set of all agents, where $n \geq 2$. Let Ω denote the non-empty and finite set of possible states.

We assume *complete information about the state* across agents. Each agent is informed of the true state $\omega \in \Omega$, while the central planner does not know it. Hence, the central planner attempts to design a mechanism that incentivizes these agents to announce about the state truthfully.

We do not assume that agents are always *selfish*, i.e., they are always concerned about their monetary interests. Agents could be *honest*, i.e., motivated not only by monetary interest but also by intrinsic preference for honesty.

We assume *incomplete information concerning honesty* in that each agent knows if she is selfish or honest, while the other agents are not informed of it. To formulate this incomplete information, we define the *type space* as follows, which is based on Bergemann and Morris (2005, 2012):

$$\Gamma \equiv (T_i, \pi_i, \theta_i)_{i \in N},$$

where $t_i \in T_i$ is agent i's type, $\theta_i : \Omega \times T_i \to \{0,1\}$, and $\pi_i : \Omega \times T_i \to \Delta(T_{-i})$.[7] Each agent i knows her type t_i as well as the state ω, but she does not know the other agents' types t_{-i}. Agent i expects that the other agents' types are distributed according to a probability measure $\pi_i(\omega, t_i) \in \Delta(T_{-i})$. Each agent is either selfish or honest: agent i is selfish (honest) if $\theta_i(\omega, t_i) = 0$ ($\theta_i(\omega, t_i) = 1$, respectively). More details will be explained later.

The central planner designs a mechanism (M, x), where $M = \underset{i \in N}{\times} M_i$, $x = (x_i)_{i \in N}$ denotes a *payment rule*, and $x_i : M \to R$ denotes the payment rule for agent i. Each agent i simultaneously announces a message $m_i \in M_i$ and obtains a monetary payment $x_i(m) \in R$ from the central planner, where we denote $m = (m_j)_{j \in N} \in M$.

We consider a class of indirect mechanisms where each agent announces a *probability distribution over states* as her message, that is,

---

[7] We denote by $\Delta(Z)$ the space of probability measures on the Borel field of a measurable space Z. We denote $Z \equiv \underset{i \in N}{\times} Z_i$, $Z_{-i} \equiv \underset{j \neq i}{\times} Z_j$, $z = (z_i)_{i \in N} \in Z$, and $z_{-i} = (z_j)_{j \neq i} \in Z_{-i}$.

$$M_i = \Delta(\Omega) \quad \text{for all} \quad i \in N .$$

We write $m_i = \omega$ if $m_i(\omega) = 1$. A *strategy* for agent i is defined as

$$s_i : \Omega \times T_i \to M_i ,$$

according to which, agent i with type t_i announces the probability distribution over states $m_i = s_i(\omega, t_i) \in \Delta(\Omega)$ when the state $\omega \in \Omega$ occurs.

If agent i is selfish, she maximizes the expected value of her utility given by monetary payment $x_i(m)$:

$$[\theta_i(\omega, t_i) = 0] \Rightarrow [\text{agent } i \text{ selects}$$

$$m_i = s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[x_i(m) \mid \omega, t_i, s_{-i}]],$$

where we assumed that the other agents announce according to $s_{-i} = (s_j)_{j \neq i}$.

On the other hand, if agent i is *honest*, she is motivated not only by monetary interest but also by an intrinsic preference for honesty, and maximizes *the expected payment minus her psychological cost*:

$$[\theta_i(\omega, t_i) = 1] \Rightarrow [\text{agent } i \text{ selects}$$

$$m_i = s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[x_i(m_i, s_{-i}(\omega, t_{-i}))$$

$$- c_i(m_i, s_{-i}(\omega, t_{-i}), \omega, t_i, G) \mid \omega, t_i]],$$

where $c_i(m, \omega, t_i, G) \in R$ denotes her psychological cost. We assume *intrinsic preference for honesty* in the manner that for every $i \in N$, $\omega \in \Omega$, $m \in M$, and $\tilde{m}_i \in M_i$,

(1) $$[\theta_i(\omega, t_i) = 1, \ m_i(\omega) > m_i'(\omega), \text{ and } x_i(\tilde{m}_i, m_{-i}) > x_i(m)]$$

$$\Rightarrow [c_i(m, \omega, t_i, G) < c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G)].$$

The assumption (1) implies that any honest agent feels more or less guilty about telling lies that generate more self-interest: hence, any honest agent strictly prefers announcing more honestly than a selfish type. In this study we allow each agent's psychological cost to be *arbitrarily small* even if this agent is honest: We do not set any condition on how much an agent cares about honesty.

An example of psychological cost is given by

$$\lambda_i\{1-m_i(\omega)\},$$

where $\lambda_i > 0$. This example describes intrinsic preference for honesty with which an agent can save psychological cost by announcing more honestly. Another example is given by

$$c_i(m,\omega,t_i,G) = \max[0, x_i(m) - x_i(\tilde{m}_i, m_{-i})]\lambda_i\{1-m_i(\omega)\},$$

where $\tilde{m}_i = \omega$ and $\lambda_i > 0$. In this example, the psychological cost depends crucially on the shape of the payment rule: the magnitude of the impact of her lie on her monetary payoff influences the size of her psychological cost. In both examples, by setting λ_i close to zero, we can consider the case in which the direct impact of the preference for honesty on an agent's decision-making can be arbitrarily small, and therefore, even honest agents are mostly motivated by their monetary interests. As implied by the latter example, we can also consider the case in which the direct impact of preference for honesty on an agent's decision-making is arbitrarily small compared with the impact of her lie on her monetary payoff.

This study investigates *Bayes Nash Equilibria* (BNE) in the game associated with a payment rule x.

3. The Theorem

We specify the payment rule $x = x^*$ as the following *quadratic scoring rule*: for every $i \in N$ and $m \in M$,

$$x_i^*(m) = -\sum_{j \neq i}[\sum_{\omega \in \Omega}\{m_i(\omega) - m_j(\omega)\}^2].$$

From a simple calculations, if s is a BNE in the game associated with x^*, then for every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

$$(2) \qquad [\theta_i(\omega,t_i) = 0] \Rightarrow [s_i(\omega,t_i) = E[\frac{\sum_{j \neq i} s_j(\omega,t_j)}{n-1} | \omega, t_i]],$$

while

$$(3) \qquad [\theta_i(\omega,t_i) = 1] \Rightarrow [\text{either } s_i(\omega,t_i)(\omega) = 1 \text{ or}$$

$$s_i(\omega,t_i)(\omega) > E[\frac{\sum_{j\neq i} s_j(\omega,t_j)(\omega)}{n-1} \mid \omega,t_i]].$$

That is, any selfish agent *mimics* the average of the other agents' announcements in expectation, while *any honest agent announces more honestly than a selfish agent*.

We define the *truthful strategy* profile s^* by

$$s_i^*(\omega,t_i) = \omega \quad \text{for all } i \in N \text{ and } (\omega,t_i) \in \Omega \times T_i,$$

according to which, each agent i announces about the state truthfully, irrespective of the state and her type. We show a necessary and sufficient condition under which the truthful strategy profile s^* is the unique BNE in the game associated with x^*, i.e., the central planner succeeds to elicit correct information about the state from the agents as unique equilibrium behavior.

We call a subset of type profiles $T \equiv \underset{i\in N}{\times} T_i$ an *event*. For convenience, for each event $E \subset T$, we write

$$\pi_i(E \mid \omega,t_i) = \pi_i(E_{-i}(t_i) \mid \omega,t_i),$$

where we denoted $E_{-i}(t_i) \equiv \{t_{-i} \in T_{-i} \mid (t_i,t_{-i}) \in E\}$. Consider an arbitrary state $\omega \in \Omega$ and an arbitrary event $E \subset T$. Let

$$V_i^1(E,\omega) \equiv \{t_i \in T_i \mid \pi_i(E \mid \omega,t_i) = 1\},$$

and

$$V_i^k(E,\omega) \equiv \{t_i \in T_i \mid \pi_i(\underset{j\in N}{\times} V_j^{k-1}(E,\omega) \mid \omega,t_i) = 1\} \quad \text{for each } k \geq 2.$$

Here, $V_i^1(E,\omega)$ implies the set of agent $i's$ types with which agent i knows that the event E and the state ω occur, and $V_i^k(E,\omega)$ implies the set of agent $i's$ types with which agent i knows that the event $\underset{j\in N}{\times} V_j^{k-1}(E,\omega)$ and the state ω occur. We then define

$$V_i^\infty(E,\omega) \equiv \bigcap_{k=1}^{\infty} V_i^k(E,\omega).$$

An event $E \subset T$ is said to be *common knowledge* at $(\omega,t) \in \Omega \times T$ if

$$t \in \underset{i\in N}{\times} V_i^\infty(E,\omega).$$

Note that if E is common knowledge at (ω,t), then

$$\pi_i(\underset{j\in N}{\times} V_j^{\infty}(E,\omega)\,|\,\omega,t_i)=1 \quad \text{for all} \quad i\in N\,.$$

We denote by $E^*(\omega)\subset T$ *the event that the state* ω *occurs and all agents are selfish*, that is,

$$E^*(\omega)\equiv\{t\in T\,|\,\forall i\in N:\theta_i(\omega,t_i)=0\}\,.$$

The Theorem: *The truthful strategy profile* s^* *is the unique BNE in the game associated with* x^* *if and only if*

$$\underset{i\in N}{\times} V_i^{\infty}(E,\omega)=\phi \quad \text{for all} \quad \omega\in\Omega.^{8}$$

By definition of common knowledge, the necessary and sufficient condition of the Theorem clearly implies that $V_i^{\infty}(E^*(\omega),\omega)=\phi$ for all $i\in N$ and $\omega\in\Omega$. The Theorem states that *all agents whether selfish or honest are willing to announce the state truthfully as unique BNE behavior, if and only if it never happens to be common knowledge that all agents are selfish*. Hence, with elimination of common knowledge of all agents' selfishness, the central planner can always succeed to elicit correct information about the state from agents. We should regard this elimination as the *minimal* requirement of an epistemological potential that an agent cares about honesty. In fact, the success of correct elicitation holds true even if it is mutual knowledge that all agents are selfish.

4. Example

For understanding the Theorem, we should capture the following characteristics of the quadratic scoring rule x^*. For simplicity of arguments, let us consider the two-agent case.

[8] By definition of common knowledge, this necessary and sufficient condition clearly implies that $V_i^{\infty}(E^*(\omega),\omega)=\phi$ for all $i\in N$ and $\omega\in\Omega$.

(a) Each agent's message space is not the set of states but the set of probability distributions over states. Hence, each agent can fine-tune her message and payment continuously.

(b) Any selfish agent is incentivized to report the same distribution as the other agent's report in expectation.

(c) Suppose that agent 1 is selfish and expects a possibility that agent 2 announces more honestly than what agent 2 expects about the announcement by agent 1 with selfish type. Then, agent 1 with selfish type has incentive to make her announcement (slightly) more honestly than what agent 2 expects about the announcement by agent 1 with selfish type. The same thing holds true even if agent 1 and agent 2 are replaced. This will be the driving force for a tail-chasing competition through which each agent announces more honestly than the other agent, reaching both agents' honest reporting.

Trivially, whenever an agent i expects a possibility that the other agent $j \neq i$ is honest, then the supposition in (c) holds and agent i is driven to be more honest. However, the other agent j does not have to be honest: it is necessary and sufficient that agent i expects a possibility that the other agent j whether selfish or honest is driven to be more honest.

To understand the logic and implication behind the Theorem, let us propose the following example with finite type space, where $n=2$, $\Omega=\{0,1\}$, $T_i=\{0,1,...,H\}$ for each $i \in \{1,2\}$. We assume that agent i is honest if and only if $t_i=0$, i.e.,

$$[\theta_i(\omega,t_i)=0] \Leftrightarrow [t_i=0].$$

Since $\Omega=\{0,1\}$, the message space of agent i is given by $M_i=[0,1]$, where $m_i \in [0,1]$ indicates the probability that the state 1 ($\omega=1$) occurs. The quadratic scoring rule is given by

$$x_1^*(m)=x_2^*(m)=-(m_1-m_2)^2.$$

We assume that there exists a common prior over type profiles π and it is symmetric, i.e., $\pi(h,h')=\pi(h',h)$ for all $(h,h') \in \{0,1,...,H\}^2$. Since the mechanism

and agents are symmetric, we often refer to an agent with type h as a "type-h agent" without specifying her identity $i \in \{1,2\}$. For simplicity of arguments, we assume that the set of selfish types $\{1,\ldots,H\}$ is (weakly) connected in the sense that

$$\pi(h,h+1) > 0 \quad \text{for all} \quad h \in \{1,...,H-1\}.$$

Without loss of generality, we assume that the true state is $\omega = 1$ (the analysis for the case of $\omega = 0$ is similar), and we drop it from the notation. The psychological cost for each agent i with honest type is given by $\lambda(1-m_i)$, where $\lambda > 0$. Let $\bar{m}_j(t_i; s_j)$ be agent $j's$ expected message conditional on agent $i's$ type t_i:

$$\bar{m}_j(t_i; s_j) \equiv E\left[s_j(t_j) \mid t_i\right] = \sum_{h=0}^{H} \pi_i(h \mid t_i) s_j(h).$$

Then, agent $i's$ best response against s_j is given by

$$BR_i(s_{-i}, t_i) = \begin{cases} \bar{m}_j(t_i; s_j) & \text{if } t_i \in \{1,\ldots,H\} \\ \min\left\{\bar{m}_j(t_i; s_j) + \lambda/2, 1\right\} & \text{if } t_i = 0 \end{cases}.$$

Hence, any honest agent is driven to be more honest than a selfish agent.

Case 1: First consider the case in which the set of selfish types is disconnected with the honest type, i.e.,

$$\pi(0,h) = 0 \quad \text{for all} \quad h \in \{1,...,H\}.$$

A selfish agent expects that the other agent is selfish with certainty, and an honest agent expects that the other agent is honest with certainty.

When $t = (0,0)$ is realized, the best response of each agent $i \in \{1,2\}$ is given by $s_i(0) = \min\{s_j(0) + \lambda/2, 1\}$, i.e., the preference for honesty drives each agent i attempt to choose a message that is slightly more honest than the other agent. Clearly, in a BNE s, $s_1(0) = s_2(0) = 1$ must be satisfied.

In contrast, an equilibrium strategy could take any value when $h \in \{1,...,H\}$. As long as there exists a constant $p \in [0,1]$ such that

$$s_i(h) = p \quad \text{for all} \quad i \in N \quad \text{and} \quad h \in \{1,\ldots,H\},$$

it is a BNE. Hence, there are infinitely many BNEs in which any selfish agents may tell a lie. Clearly, we fail to elicit the correct state as a unique BNE in Case 1.

Case 2: Next consider the case in which, unlike Case 1, the set of selfish types is connected with the honest type in such a minimal sense that there exists $h \in \{1, 2, ..., H\}$ such that $\pi(0, h) > 0$. For simplicity, we assume $h = 1$, i.e.,

$$\pi(0,1) > 0.$$

It will be easy to see that the same argument holds true even if we replace 1 with any $h \in \{2, ..., H\}$.

Due to selfish agents' higher-order reasoning, this minimal connection drastically changes the set of BNEs as follows. Clearly, a type-0 (honest) agent is driven to be more honest because of her intrinsic preference for honesty. The minimal connection implies that a type-1 agent expects that the other agent may be type-0 with a positive probability and she would like to match her message with the other agent's announcement in expectation. Hence, the type-1 agent is also driven to be more honest. Similarly, a type-2 agent expects that the other agent may be type-1 with a positive probability, and therefore, is driven to be more honest. We can iterate this argument and verify that any agent whether selfish or honest is driven to be more honest; i.e., attempts to send a more honest message than the other agent. This structure of best responses immediately leads us to the uniqueness of BNE, where all agents whether selfish or honest report $m_i = \omega = 1$ truthfully.

Note that this uniqueness holds even if both agents' selfishness is mutual knowledge. As long as $t_1 \geq 2$ and $t_2 \geq 2$, each agent does not expect that the other agent may be honest. However, the above mentioned higher-order reasoning will guide any selfish agent to send a more truthful message, which drastically shrinks the set of BNEs. As long as there is no common knowledge of both agents' selfishness, this logic always functions and the uniqueness of BNE is guaranteed.

Case 1 corresponds to situations in which all selfish types completely eliminate association with honest types, meeting a failure of unique information elicitation. However, this failure is exceptional. If there is at least one selfish type who expects an

(possibly indirect) influence of an honest type even a little, then unique information elicitation is achievable. The driving force behind this is not that more people become honest, but simply that *selfish people do not rule out the existence of honest agents from their considerations.*

5. Proof of The Theorem

It is clear from (2) and (3) that s^* is a BNE. Suppose that s is a BNE. Fix an arbitrary $\omega \in \Omega$. Let

$$\alpha \equiv \min_{(i,t_i)} s_i(\omega, t_i)(\omega),$$

and

$$\tilde{T}_i \equiv \{t_i \in T_i \mid s_i(\omega, t_i)(\omega) = \alpha\} \quad \text{for each } i \in N.$$

Suppose that

$$\times_{i \in N} V_i^{\infty}(E, \omega) = \phi \quad \text{for all } \omega \in \Omega.$$

From the definition of common knowledge, this supposition implies that

$$V_i^{\infty}(E^*(\omega), \omega) = \phi \quad \text{for all } i \in N \text{ and } \omega \in \Omega.$$

Towards a contradiction, suppose that

$$\alpha < 1,$$

which implies that there exists a type who is selfish and announces dishonestly. Note from (2) and (3) that any honest agent prefers announcing more honestly than selfish agents, implying that no honest type belongs to $\tilde{T}_i$:

$$[t_i \in \tilde{T}_i] \Rightarrow [\theta_i(\omega, t_i) = 0].$$

Consider an arbitrary $i \in N$ and $t_i \in \tilde{T}_i$. From (2) and (3), α is equal to the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, type t_i expects that any other agent $j \neq i$ announces $m_j(\omega) = \alpha$, that is,

$$\pi_i(\times_{j \in N} \tilde{T}_j \mid \omega, t_i) = 1.$$

This along with (2) and (3) implies that agent i with type t_i expects that the other agents are never honest, that is,

$$\pi_i(E^*(\omega) \mid \omega, t_i) = 1 .$$

Hence, we have

$$\tilde{T}_i \subset V_i^1(E^*(\omega), \omega) .$$

Moreover, since

$$\pi_i(\underset{j \in N}{\times} V_i^1(E^*(\omega), \omega) \mid \omega, t_i) \geq \pi_i(\underset{j \in N}{\times} \tilde{T}_i \mid \omega, t_i) = 1,$$

we have $\pi_i(\underset{j \in N}{\times} V_i^1(E^*(\omega), \omega) \mid \omega, t_i) = 1$, that is,

$$\tilde{T}_i \subset V_i^2(E^*(\omega), \omega) .$$

Similarly, we have

$$\tilde{T}_i \subset V_i^k(E^*(\omega), \omega) \quad \text{for all} \quad k \geq 2 .$$

Hence, we have

$$\tilde{T}_i \subset V_i^\infty(E^*(\omega), \omega) ,$$

which however contradicts the supposition that $V_i^\infty(E^*(\omega), \omega) = \phi$. Hence, we conclude $\alpha = 1$, that is, $s = s^*$, and therefore, we have proved the "if" part of the Theorem.

Fix an arbitrary $\omega' \neq \omega$. We specify a strategy profile s^+ as follows: for every $i \in N$ and $t_i \in T_i$,

$$s_i^+(\omega, t_i) = \omega \qquad \text{if} \quad t_i \notin V_i^\infty(E^*(\omega), \omega) ,$$

$$s_i^+(\omega, t_i) = \omega' \qquad \text{if} \quad t_i \in V_i^\infty(E^*(\omega), \omega) ,$$

and

$$s_i^+(\tilde{\omega}, t_i) = s_i^*(\tilde{\omega}, t_i) \quad \text{for all} \quad \tilde{\omega} \neq \omega .$$

It is clear from (2) and the above argument that s^+ is a BNE, and $s^+ \neq s^*$ whenever $V_i^\infty(E^*(\omega), \omega) \neq \phi$ for some $i \in N$. Hence, we have proved the "only-if" part of the Theorem.

Remark 1: This study has considered only pure strategy BNE. However, we can directly use the same logic to the uniqueness of *mixed strategy* BNE. Because of quadratic scoring rule, irrespective of whether the other agents' strategies are mixed or pure, any selfish

agent prefers announcing the same distribution as the other agents' announcements in expectation, while any honest agent prefers announcing more honestly than a selfish agent: the resultant tail-chasing competition eliminates any unwanted BNE, whether it is pure or mixed.

Remark 2: The specified payment rule x^* does not satisfy budget-balancing. In fact, in the two-agent case, it is hard to find an alternative rule that induces unique information elicitation like x^* and also satisfies budget-balancing. On the other hand, in the three-or-more-agent case, it is easy to check that the following payment rule induces unique information elicitation and satisfies budget-balancing: for every $i \in N$ and $m \in M$,

$$x_i^+(m) = x_i^*(m) + r_i(m_{-i}),$$

where

$$r_i(m_{-i}) \equiv \frac{1}{n-2} \sum_{i' \neq i, j \neq i, i' \neq j} [\sum_{\omega \in \Omega} \{m_{i'}(\omega) - m_j(\omega)\}^2].$$

Remark 3: The Theorem holds true irrespective of the number of agents participating in the information elicitation problem. However, the restrictiveness of the necessary and sufficient condition depends crucially on the number of participants. Namely, the more agents participate in the central planner's problem, the less likely it is to be common knowledge that all agents are selfish.

Moreover, even if the number of participants is limited, the central planner should recruit informed people from a wider range. Let us go back to the example in Section 4. Suppose that the central planner picks up agent 1 from a narrower range than what Case 2 assumes, eliminating the possibility of type 0 to be chosen. Then, the set of selfish types becomes disconnected with the honest type: unique information elicitation fails.

6. Discussion

6. 1. Unique Implementation of Social Choice Function

This study did not explicitly consider what is the purpose of the central planner to elicit correct information about the state from agents. If the central planner aims to determine a desirable resource allocation in a contingent manner on the state, each agent's self-interest motive may be influenced not only by the monetary payment but also by this allocation determination. Hence, the central planner must design a mechanism as a combination of a payment rule and an allocation rule to incentivize agents to announce about the state sincerely as unique equilibrium behavior and uniquely implement a social choice function that describes the central planner's desirable state-contingent allocation. That is, the central planner must solve the unique implementation problem of a social choice function.

If preferences are quasi-linear and large transfers are permitted, we can solve the unique implementation problem by almost directly applying this study's theorem: large transfers negate the effects of allocation determination.

More importantly, a companion work by Matsushima (2020b) extended the Theorem to this unique implementation problem without assuming quasi-linearity and with using only small monetary transfers. Matsushima decomposed mechanism design into two parts: the first part corresponds to unique information elicitation implied by this study, and the second part corresponds to unique implementation with provability implied by Matsushima (2020a). By using the Theorem for the first part, Matsushima (2020b) could show a very permissive result that any social choice function is uniquely implementable in BNE whenever it never happens to be common knowledge that all agents are selfish.

6. 2. Smart Contracts

Implementation theory generally assumes that there exists a central planner who has a power to force allocations and payments according to the predetermined mechanism or contracts. This study has followed the assumption that such a central planner exists for convenience of arguments. However, when considering the social implementation of this study's result, the Theorem essentially does not require the presence of such a central planner.

Another companion work by Matsushima and Noda (2020) pointed out that the mechanism constructed in this study can be socially implemented within the scope of current digital technology, by making payments in *digital currencies*, programming the mechanism as a *smart contract*, and managing it on a *blockchain* network such as Ethereum. Matsushima and Noda (2020) showed that the information elicitation mechanism proposed in this paper acts as a *digital court*: the information elicitation mechanism will punish any deviation from the pre-agreed contract in absence of a central planner and trusted third parties.

6. 3. Asymmetric Information

This study has assumed symmetric information about the state: multiple agents share the same information. It is an important extension of this study to investigate the *asymmetric information* environment where agents can access their respective private information channels and the central planner wants to elicit correct private information from every agent. For a more discussion, see Appendix A, which demonstrates an example that expresses some of the difficulty in this extension.

6. 4. Uncertainty in Information Access

This study has assumed that all agents can certainly observe the true state. We can eliminate this assumption with no substantial change of this study's argument: we can directly apply the basic logic behind the Theorem to uncertain environments in which an agent fails to access information channel. Appendix B gives a more precise argument on this issue. Note that there is another benefit of having a large number of people participate in the central planner's problem, because it increases the chances that someone will have access to the information: the more people participate, the more certainly the central planner successfully elicits information from participants.

7. Conclusion

This study investigated the unique information elicitation problem under complete information about the state. We permitted the central planner to use only payment rule design without public monitoring technology. We assumed that it never happens to be common knowledge that all agents are selfish, and showed a very permissive result that the central planner can elicit correct information about the state from agents as their unique BNE behavior. This result indicates that a potential of social implementation of information elicitation devices are much greater than what a standard model with all agents' selfishness has expected.

It is the most important future research to investigate the environments with asymmetric information concerning the state. Can quadratic scoring rules still function? If not, what is the design of payment rule that solves unique information elicitation? How do we define intrinsic preference for honesty in this environment? These questions are just the tip of the iceberg in future research, but all of them could include new theoretical substances beyond this study's scope.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for Truth-Telling," *Econometrica* 87, 1115–1153.

Aoyagi, M. (1998): "Correlated Types and Bayesian Incentive Compatible Mechanisms with Budget Balance," *Journal of Economic Theory* 79, 142–151.

Bergemann, D. and S. Morris (2005): "Robust mechanism design," *Econometrica* **73**, 1771–1813.

Bergemann, D. and S. Morris (2012): "An Introduction to Robust Mechanism Design," *Foundations and Trends in Microeconomics* 8 (3), 169–230.

Brier, G. (1950): "Verification of Forecasts Expressed in Terms of Probability", *Monthly Weather Review* 78, 1–3.

Carlsson, H. and E. van Damme (1993): "Global Games and Equilibrium Selection," *Econometrica* 61, 989–1018.

Cooke, R. (1991): *Experts in Uncertainty: Opinion and Subjective Probability in Science,* New York: Oxford University Press.

Charness, G. and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica* 76 (6), 1579-1601.

Dasgupta, A. and A. Ghosh (2013): "Crowdsourced Judgement Elicitation with Endogenous Proficiency," *In Proceedings of the 22nd International Conference on World Wide Web*, 319-330.

Dutta, B. and A. Sen (2012): "Nash Implementation with Partially Honest Individuals," Games and Economic Behavior 74 (1), 154-169.

Ellingsen, T. and M. Johannesson (2004): "Promises, Threats and Fairness," *The Economic Journal* 114 (495), 397-420.

Johnson, S., J. Pratt, and R. Zeckhauser (1990): "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica* 58, 873–900.

Kartik, N. (2009): "Strategic Communication with Lying Costs," *The Review of Economic Studies* 76 (4), 1359-1395.

Kartik, N., M. Ottaviani, and F. Squintani (2007): "Credulity, Lies, and Costly Talk," *Journal of Economic Theory* 134 (1), 93–116.

Kartik, N., and O. Tercieux (2012): "Implementation with Evidence," *Theoretical Economics* 7 (2), 323-355.

Kartik, N., O. Tercieux, and R. Holden. (2014): "Simple Mechanisms and Preferences for Honesty," *Games and Economic Behavior* 83, 284-290.

Kong, Y. and G. Shoenebeck (2019): "An Information Theoretic Framework for Designing Information Elicitation Mechanisms that Reward Truth-Telling," *ACM Transactions on Economics and Computation (TEAC),* 7 (1), 1-33.

Koszegi, B. (2014): "Behavioral Contract Theory," *Journal of Economic Literature* 52 (4), 1075-1118.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982): "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic theory* 27, 245–252.

Krishna, V. (2009): *Auction Theory*, Academic press.

Maskin, E., and T. Sjöström (2002): "Implementation Theory," in: K. Arrow, A. Sen, and K. Suzumura (Eds.) *Handbook of Social Choice and Welfare Volume 1*, Elsevier.

Matsushima, H. (1990): "Dominant Strategy Mechanisms with Mutually Payoff-Relevant Information and with Public Information," *Economics Letters* 34, 109–112.

Matsushima, H. (1991): "Incentive Compatible Mechanisms with Full Transferability," *Journal of Economic Theory* 54, 198-203.

Matsushima, H. (1993): "Bayesian Monotonicity with Side Payments," *Journal of Economic Theory* 59, 107–121.

Matsushima, H. (2007): "Mechanism Design with Side Payments: Individual Rationality and Iterative Dominance," *Journal of Economic Theory* 133 (1), 1-30.

Matsushima, H. (2008a): "Role of Honesty in Full Implementation," *Journal of Economic Theory* 139, 353–359.

Matsushima, H. (2008b): "Behavioral Aspects of Implementation Theory," *Economics Letters* 100(1), 161-164.

Matsushima, H. (2020a): "Implementation without Expected Utility: Ex-Post Verifiability," *Social Choice and Welfare* 53 (4), 575-585.

Matsushima, H. (2020b): "Implementation, Honesty and Common Knowledge," mimeograph.

Matsushima, H. and S. Noda (2020): "Mechanism Design with Blockchain Enforcement,"

CARF-F-474, University of Tokyo.

Mazar, N., O. Amir, and D. Ariely (2008): “More Ways to Cheat-Expanding the Scope of Dishonesty,” *Journal of Marketing Research* 45 (6), 651-653.

Miller, N., J. Pratt, R. Zeckhauser, and S. Johnson (2007): “Mechanism Design with Multidimensional, Continuous Types and Interdependent Valuations,” *Journal of Economic Theory* 136 (1), 476-496.

Miller, N., P. Resnick, and R. Zeckhauser (2005): “Eliciting Informative Feedback: The Peer-Prediction Method,” *Management Science* 51 (9), 1359-1373.

Morris, S., and H. S. Shin (1998): “Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks,” *American Economic Review* 88 (3), 587–597.

Palfrey, T. R. (2002): Implementation theory, *Handbook of Game Theory with Economic Applications* 3, 2271-2326.

Postlewaite, A. and X. Vives (1987): “Bank runs as an equilibrium phenomenon,” *Journal of Political Economy* 95, 485–491.

Prelec, D. (2004): “A Bayesian Truth Serum for Subjective Data,” *Science*, 306 (5695), 462-466.

Salanié, B. (1997): *The Economics of Contracts: A Primer*, Cambridge, MA: MIT Press.

Appendix A: Asymmetric Information

Let us consider an example with finite type space: $n=2$, $\Omega=\Omega_1\times\Omega_2=\{0,1\}^2$, and $T_i=\{0,1\}$, where each agent $i\in\{0,1\}$ privately observes $\omega_i\in\{0,1\}$ with equal probability, and $t_i=0$ ($t_i=1$) implies that agent i is selfish (honest, respectively). Each agent i announces a message $m_i\in[0,1]$, implying the probability of $\omega_i=0$. Consider the following payment rule:

$$x_1^*(m)=x_2^*(m)=-(m_1-m_2)^2.$$

For simplicity, we assume that any honest type of each agent i announces $m_i=\omega_i$, while any selfish type maximizes the expected payment. Each agent $i's$ expectation about the other agent $j's$ private information and type is described by $\pi_i(\omega_j,t_j\mid\omega_i,t_i)$, which implies the probability that the other agent j observes private information ω_j and has type t_j provided that payer i observes private information ω_i and has type t_i.

Assume that there exist $q^0>0$ and $q^1>0$ such that

$$q^0=\pi_1(0,1\mid 0,0)=\pi_1(0,1\mid 1,0)=\pi_2(0,1\mid 0,0)=\pi_2(0,1\mid 1,0),$$

and

$$q^1=\pi_1(1,1\mid 0,0)=\pi_1(1,1\mid 1,0)=\pi_2(1,1\mid 0,0)=\pi_2(1,1\mid 1,0).$$

This assumption implies that whenever each payer's private information is perfectly correlated, i.e., in the complete information environment about the state. This assumption eliminates the common knowledge of all agents' selfishness.

Let

$$p=\frac{q^1}{q^0+q^1}.$$

We can show that any selfish agent $i's$ announcing $m_i=p$ regardless of the realization of ω_i is a BNE. Hence, the central planner fails to elicit correct information from selfish agents, even if the common knowledge of all agents' selfishness is eliminated.

Appendix B: Uncertainty in Information Access

Appendix B considers the case in which the central planner does not know if each agent can access the information channel. We modify the type space as follows:

$$\Gamma \equiv (T_i, \pi_i, \theta_i, \eta_i)_{i \in N}, \text{ where } \eta_i : T_i \to \{0,1\}.$$

Agent i is informed (uninformed) if $\eta_i(t_i) = 0$ ($\eta_i(t_i) = 1$, respectively). Agent i with $\eta_i(t_i) = 0$ ($\eta_i(t_i) = 1$) can (cannot, respectively) access the information channel and observe the state. We assume that any agent always expects that there exist other agents who are informed with a positive probability:

Assumption 1: For every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

$$\pi_i(\{t_{-i} \in T_{-i} \mid \theta_j(\omega, t_j) = 0 \text{ for some } j \neq i\} \mid \omega, t_i) > 0.$$

We also assume that any uninformed agent is honest: hence, we categorize agents into three cases, i.e., "selfish and informed," "honest and informed," and "honest and uninformed."[9]

Assumption 2: For every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

$$[\pi_i(t_i) = 1] \Rightarrow [\theta_i(\omega, t_i) = 0].$$

We consider a class of mechanisms (M, x) where each agent announces either a probability distribution on Ω or "μ":

$$M_i = \Delta(\Omega) \bigcup \{\mu\} \text{ for all } i \in N.$$

The announcement of "μ" implies that she is uninformed.

If agent i is selfish and informed, she maximizes the expected payment:

$$[\theta_i(\omega, t_i) = 0 \text{ and } \eta_i(t_i) = 0]$$

[9] Consideration of cases without this assumption may be carefully discussed in future work, though the benefits of lying to be an uninformed agent are generally negative.

$$\Rightarrow \quad [\text{agent } i \text{ selects}$$

$$m_i = s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[x_i(m_i, s_{-i}(\omega, t_{-i})) \mid \omega, t_i]].$$

If agent i is honest and informed, she maximizes the expected payment minus her psychological cost:

$$[\theta_i(\omega, t_i) = 1 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \quad [\text{agent } i \text{ selects } m_i = s_i(\omega, t_i)$$

$$\in \arg\max_{m_i \in M_i} E[x_i(m_i, s_{-i}(\omega, t_{-i})) - c_i(m_i, \omega, t_i) \mid \omega, t_i]].$$

We assume that

$$[\theta_i(\omega, t_i) = 1, \ \eta_i(t_i) = 0, \text{ and } m_i(\omega) > m_i'(\omega)]$$

$$\Rightarrow \quad [c_i(m_i, \omega, t_i) < c_i(m_i', \omega, t_i)],$$

and

$$[\theta_i(\omega, t_i) = 1 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \quad [c_i(\mu, \omega, t_i) \geq c_i(m_i, \omega, t_i) \text{ for all } m_i \neq \mu].$$

With the latter inequalities, pretending to be uninformed is more dishonest than announcing about the state incorrectly. For simplicity of arguments, we assume that if agent i is honest and uninformed, she announces μ:

$$[\theta_i(\omega, t_i) = 1 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \quad [\text{agent } i \text{ selects } m_i = s_i(\omega, t_i) = \mu].$$

We specify the payment rule x^{**} as a modification of x^{*}: for every $i \in N$ and $m \in M$,

$$x_i^{**}(m) = -\sum_{\substack{j \neq i \\ m_j \neq \mu}} [\sum_{\omega \in \Omega} \{m_i(\omega) - m_j(\omega)\}^2]$$

$$\text{if } m_i \in \Delta(\Omega),$$

and

$$x_i^{**}(m) = -\varepsilon - \max_{\tilde{m}_i \in \Delta(\Omega)} \sum_{\substack{j \neq i \\ m_j \in \Delta(\Omega)}} [\sum_{\omega \in \Omega} \{\tilde{m}_i(\omega) - m_j(\omega)\}^2]$$

$$\text{if } m_i = \mu,$$

where $\varepsilon > 0$. Note that

$$\min_{m_i \in \Delta(\Omega)} x_i^{**}(m) = x_i^{**}(\mu, m_{-i}) + \varepsilon \quad \text{for all} \quad m_{-i} \in M_{-i}.$$

From these specifications and assumptions, if s is a BNE in the game associated with x^{**}, then, for every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

(B-1) $$[\theta_i(\omega, t_i) = 0 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \quad [s_i(\omega, t_i) = E[\frac{\sum_{j \neq i, m_j \neq \mu} s_j(\omega, t_j)}{|\{j \neq i \mid m_j \neq 0\}|} \mid \omega, t_i]],$$

(B-2) $$[\theta_i(\omega, t_i) = 1 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \quad [\text{either } s_i(\omega, t_i)(\omega) = 1 \text{ or}$$

$$s_i(\omega, t_i)(\omega) > E[\frac{\sum_{j \neq i, m_j \neq \mu} s_j(\omega, t_j)(\omega)}{|\{j \neq i \mid m_j \neq 0\}|} \mid \omega, t_i]],$$

and

$$[\theta_i(\omega, t_i) = 1 \text{ and } \eta_i(t_i) = 1]$$

$$\Rightarrow \quad [s_i(\omega, t_i) = \mu].$$

Any selfish and informed agent mimics the average of the other informed agents' announcements in expectation. Any honest and informed agent announces more honestly than a selfish agent. Any honest and uninformed agent truthfully reports the fact that she is uninformed.

We define the truthful strategy profile s^{**} as follows: for every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

$$s_i^{**}(\omega, t_i)(\omega) = 1 \qquad \text{if } \eta_i(t_i) = 0,$$

and

$$s_i^{**}(\omega, t_i) = \mu \qquad \text{if } \eta_i(t_i) = 1.$$

We denote by $E^{**}(\omega) \subset T$ the event that the state ω occurs and there exists no agent who is honest and informed, that is,

$$E^{**}(\omega) \equiv \{t \in T \mid \forall i \in N : (\theta_i(\omega, t_i), \eta_i(t_i)) \neq (1, 0)\}.$$

Theorem B: *The truthful strategy profile* s^{**} *is the unique BNE in the game associated with* x^{**} *if and only if*

$$\underset{i\in N}{\times} V_i^{\infty}(E^{**}(\omega),\omega)=\phi \text{ for all } \omega\in\Omega.$$

Proof: It is clear that s^{**} is a BNE. Suppose that s is a BNE. Fix an arbitrary $\omega\in\Omega$. Let

$$\alpha \equiv \min_{(i,t_i),\eta_i(t_i)=0} s_i(\omega,t_i)(\omega),$$

and

$$\tilde{T}_i \equiv \{t_i \in T_i \mid s_i(\omega,t_i)(\omega)=\alpha\} \text{ for each } i\in N.$$

Suppose that $V_i^{\infty}(E^{**}(\omega),\omega)=\phi$ for all $i\in N$, and

$$\alpha<1,$$

which implies that there exists a type who is informed and announces dishonestly. Note

$$[t_i\in\tilde{T}_i]\Rightarrow[\theta_i(\omega,t_i)=0].$$

Consider an arbitrary $i\in N$ and $t_i\in\tilde{T}_i$. From (B-1) and (B-2), α is equal to the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, agent i expects that any other informed agent $j\neq i$ announces $m_j(\omega)=\alpha$, that is,

$$\pi_i(\underset{j\in N}{\times}(\tilde{T}_j\cup\hat{T}_j)\mid\omega,t_i)=1,$$

where $\hat{T}_j$ denotes the set of agent $j's$ types which are uninformed, that is,

$$\hat{T}_j\equiv\{t_j\in T_j\mid\eta_j(t_j)=1\}.$$

Since $\underset{j\in N}{\times}(\tilde{T}_j\cup\hat{T}_j)\subset E^{**}(\omega)$, agent i expects the other agents are either selfish or uninformed, that is,

$$\pi_i(E^{**}(\omega)\mid\omega,t_i)=1.$$

Hence, we have

$$\underset{j\in N}{\times}(\tilde{T}_j\cup\hat{T}_j)\subset V_i^1(E^{**}(\omega),\omega).$$

Moreover, since

$$\pi_i(\underset{j\in N}{\times} V_i^1(E^{**}(\omega),\omega)\,|\,\omega,t_i) \geq \pi_i(\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j)\,|\,\omega,t_i) = 1,$$

we have $\pi_i(\underset{j\in N}{\times} V_i^1(E^{**}(\omega),\omega)\,|\,\omega,t_i) = 1$, that is,

$$\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j) \subset V_i^2(E^{**}(\omega),\omega).$$

Similarly, we have

$$\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j) \subset V_i^k(E^{**}(\omega),\omega) \quad \text{for all} \quad k \geq 2.$$

Hence, we have

$$\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j) \subset V_i^\infty(E^{**}(\omega),\omega),$$

which however contradicts the supposition that $V_i^\infty(E^{**}(\omega),\omega) = \phi$. Hence, we conclude $\alpha = 1$, that is, $s = s^{**}$, and therefore, we have proved the "if" part.

Q.E.D.