CENTER FOR ADVANCED RESEARCH IN FINANCE

CARF Working Paper

CARF-F-549

Social interaction and epistemology in information elicitation

Hitoshi Matsushima
Department of Economics, University of Tokyo

December 7, 2022

CARF is presently supported by Nomura Holdings, Inc., Sumitomo Mitsui Banking Corporation, The Dai-ichi Life Insurance Company, Limited, The Norinchukin Bank, MUFG Bank, Ltd. and Ernst & Young ShinNihon LLC. This financial support enables us to issue CARF Working Papers.

Social interaction and epistemology in information elicitation

Hitoshi Matsushima

Department of Economics, University of Tokyo,
Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Email: hitoshi@e.u-tokyo.ac.jp

December 7, 2022

Social interaction and epistemology in information elicitation

Abstract

We consider the possibility that in a society where innately prosocial and adversarial agents exist albeit in the minority, the majority of agents behave honestly from two distinct perspectives. First, we consider a socioeconomic perspective in which the majority are influenced by their partners through iterative social interaction with conformity. We define an honest society as one in which the majority can acquire the prosocial mode after such an iteration. Second, we consider a strategic perspective in which the majority are motivated by self-interest and use epistemological reasoning. We show an equivalence between these perspectives in information elicitation, implying that a society is honest if and only if the majority behaves honestly via a unique Bayesian Nash equilibrium in a detail-free mechanism.

Keywords: adversarial agents, conformity, social network, honest society, detail-free mechanism, uniqueness.
JEL Classification: C72, D71, D78, H41

1. Introduction

Whether people are motivated to behave honestly is a central question in economics and social science. A situation that makes this question intractable arises when a central planner lacks relevant information and informed people are not selfishly interested in the central planner's intentions. This is particularly the case when the central planner lacks detailed ethical information on the fair distribution. Informed individuals are rarely inherently prosocial. Conversely, it is conceivable that they have innately adversarial (anti-social and spiteful) tendencies. Furthermore, there is a concern that many ordinary

people will behave dishonestly under the influence of such bad people. Hence, we are pressured to construct a new incentive theory for the successful collection of hidden information.

Based on the awareness of the above issues, we consider the possibility that the majority of agents in a society behave honestly from two distinct perspectives—a socioeconomic perspective and a strategic one—and relate these perspectives by showing some equivalence theorems in the context of information elicitation. We see such a society as a place where two agents are randomly matched into partnerships. Each agent is prosocial, adversarial, or neutral. A prosocial agent has an innate mastery of the prosocial mode and behaves honestly, whereas an adversarial agent has an innate mastery of the adversarial mode and behaves dishonestly.

Neutral agents, by contrast, who make up the majority of the society, do not have any such innate behavioral modes. However, if prosocial agents could exist, neutral agents may learn to behave honestly through interdependence in a social network, directly through social interaction, or indirectly through epistemological reasoning. Conversely, if adversarial agents are present, neutral agents may learn to behave dishonestly. This study elucidates the conditions under which neutral agents would behave honestly without any innate behavioral modes, provided that both prosocial and adversarial agents could exist in the society, albeit in a minority.

First, from a socioeconomic perspective, we consider a model in which, owing to conformity, a neutral agent is influenced by the partners' behavioral mode through iterative social interaction. A neutral agent follows the prosocial (adversarial) behavioral mode and behaves honestly (dishonestly) if the partners are more (less) likely to behave honestly than dishonestly.

We call a society stable if each neutral agent eventually masters some behavioral mode, whether prosocial or adversarial, through social interactions. We call a society honest (dishonest) if all neutral agents eventually master the prosocial (adversarial) behavioral mode. If a society is honest and social interaction with conformity functions ideally, there is a high probability that a central planner who lacks information about the state that is needed to make desirable decisions can elicit the correct information from

informed agents by listening directly to them, without having to design any artificial incentive.

Second, from a strategic perspective, we investigate a strategic model that differs from the socioeconomic model, where social interaction does not function, neutral agents never master particular behavioral modes, and they are instead concerned with nothing but the selfish maximization of their monetary interests. In this model, the central planner must design artificial incentive devices that motivate neutral agents through epistemological reasoning.

Specifically, the central planner asks both the agents of a partnership the same question several times, such as which state is likely to occur. If an agent does not change the answer in the end, the central planner considers this answer as the agent's true intention. Before this procedure, the central planner sets up a side-payment rule (i.e., a mechanism) that is designed using a quadratic scoring rule (Brier, 1950) and its variant, according to which the central planner compares the answers of the two agents and gives them monetary penalties if their answers differ.

The central planner must design a mechanism with two points in mind: detail freeness and uniqueness. The mechanism must be detail-free, that is, it must be independent of the fine detail of the specifications of the society, because the central planner is unaware of it. The mechanism must satisfy uniqueness, that is, it must uniquely determine all the agents' behaviors. In other words, a game associated with the society and mechanism must have a unique Bayesian Nash equilibrium (BNE).

We call a mechanism solvable in a society if the associated game has a unique BNE, and this equilibrium brings the central planner to converged answers, whether correct or incorrect. We then show an equivalence theorem between the socioeconomic model and the strategic model in that there exists a detail-free mechanism such that a society is stable if and only if it makes this mechanism solvable. This mechanism motivates neutral agents to make their final answers the same as the answers they make in accordance with the behavioral mode acquired through iterative social interaction. Importantly, we show another equivalence that the central planner can elicit the correct information after iterative social interaction (i.e., a society is honest) in the socioeconomic model if and only if the central planner can the elicit correct information via a unique BNE in the strategic model.

This study captures the foundation of a society as a social network defined based on the distribution of partnership-type profiles. The socioeconomic perspective interprets this network as a stochastic pattern of matching through iterative social interaction, whereas the strategic perspective interprets it as descriptions of agents' epistemology regarding randomly matched partners' types. The novelty of this study lies in stating that, although these perspectives correspond to their respective behavioral models, their possibilities for solving the same information elicitation problem are deeply related, as our equivalence theorems indicate.

The socioeconomic model is related to the social theory literature that studies the effects of social influence, social learning, and social contagion on well-being and performance.[1] This study departs from the literature in that it considers the presence of pre-fixed prosocial and adversarial types. The social theory literature exclusively analyzes the impacts of tie strength and network shape, such as bonding and bridging, on well-being and performance. This study indicates that these impacts can be reversed as positive or negative, depending on where pre-fixed prosocial and adversarial types are positioned in the social network.

The strategic model is related to the behavioral implementation theory explored by Matsushima (2008a; 2008b), which investigates the possibility of a mechanism design that implements a social choice function via unique equilibrium behavior, provided agents are either honest or selfish (neutral), but not adversarial.[2] Notably, Matsushima (2022a) shows that with three or more agents, under complete information about the state, any social choice function is uniquely implementable in a BNE if "all the agents are selfish" is never common knowledge; that is, if no common knowledge of selfishness (i.e., NCKS) holds. In this study, we focus on a simple information elicitation problem and extend behavioral implementation theory by considering adversarial agents in addition to prosocial agents and neutral (selfish) agents. We generalize these positive results by replacing NCKS with the honesty of society, which is less restrictive than NCKS.

[1] See, for example, Granovetter (1973), Coleman (1988), Kohler et al. (2001), Putnam (2000, 2004), Christakis and Fowler (2007), Rosenquist et al. (2011), and Burt (2018.)
[2] See, for example, Dutta and Sen (2012), Matsushima (2013), Kartik et al. (2014), Saporiti (2014), Ortner (2015), and Mukherjee et al. (2017).

In the presence of adversarial types, we need a different mechanism design from that of Matsushima (2022a). The technical novelty of this study lies in devising a new mechanism design to eliminate adversarial agents' influence on neutral agents' epistemological consideration. In Matsushima (2022a), owing to the non-existence of adversarial types, it suffices for the central planner to ask an agent for a single answer. However, to eliminate the strategic impacts of adversarial agents on neutral agents' epistemological consideration, we introduce a new design method according to which the central planner asks the same questions to each agent repeatedly and adopts the final answers as their true intention.

Abeler et al. (2019) show that subjects who are in trade-offs between material interest and honesty give up a large proportion of potential benefits from lying, whereas malevolent motivations are largely drowned out by monetary interests. Nevertheless, the extensions made in this study are of great value. We find that even a small possibility of adversarial types compared with prosocial types may make all the differences, depending on where these types are positioned in the social network. Conversely, even in a society where adversarial motives are prevalent, it is possible to explain how a slight possibility of the existence of prosocial agents can lead the society in a positive direction. This study opens up a new future research direction on incentives that is relevant to network formation (Bala and Goyal, 2000) and social mobility (Chetty et al, 2022a, 2022b).

People interact with others in various locations such as homes, schools, workplaces, local communities, and social networking sites. People repeatedly learn from and are influenced by other people and the pattern of such repeated contacts is stable. The appearance of this pattern is intrinsically related to solving the central question in economics and social science (such as incentive issues for fairness concerns).

The socioeconomic perspective assumes that people tend to be emotionally affected by and conform to the feelings and attitudes of others, which aids their problem-solving. In other words, collectivism may enable many people to acquire prosocial modes through social interaction. However, we need to be careful because, if society is not decent, the presence of only a handful of psychopaths can turn many ordinary citizens into sociopaths.

By contrast, the strategic perspective sees people as cold-blooded beings who are not subject to the mental influence of others and instead use their cold brains to recognize the pattern in an individualistic manner. The epistemological knowledge of the pattern

combined with the mechanism design by the central planner can contribute to solving the problem. Such individualism can thus achieve the same level of prosociality as achieved by collectivism, but no more.

In a social environment with social common capital (Uzawa, 2005) and relational social capital (Putnam, 2000, 2004), collectivism provides solutions to many problems in a decent society. However, in a social environment that inadequately manages social common capital and relational social capital, even decent societies must resort to individualistic approaches using artificial methods such as financial incentives.

This study does not ask which of the two perspectives is superior or which model is appropriate as a depiction of reality. Rather, it shows that the ambiguity in the interpretation of social networks allows the two perspectives to be theoretically related and comparatively examined.

The remainder of this paper is organized as follows. In Section 2, we explain the partnerships and behavioral modes. In Section 3, we demonstrate the socioeconomic model, explain iterative social interaction with conformity, and define a stable society. In Section 4, we define an honest society and investigate examples to understand the case in which a society is honest. In Section 5, we demonstrate the strategic model, investigate the information elicitation problem, and present equivalence theorems. Section 6 concludes the paper.

2. Partnerships and behavioral modes

Consider a randomly matched partnership consisting of agents 1 and 2. We define a society as a type space (T, π, θ), where $T = T_1 \times T_2$, T_i denotes a finite set of agent i's types, $\pi : T_1 \times T_2 \to [0,1]$ denotes a distribution over type profiles $(t_1, t_2) \in T_1 \times T_2$, $\theta = (\theta_1, \theta_2)$, and $\theta_i : T_i \to \{P, A, N\}$. Each agent $i \in \{1,2\}$ is prosocial (P), adversarial (A), or neutral (N), which we denote as $\theta_i(t_i) \in \{P, A, N\}$, depending on their type $t_i \in T_i$. A prosocial agent (i.e., $\theta_i(t_i) = P$) always behaves honestly, whereas an adversarial agent (i.e., $\theta_i(t_i) = A$) always behaves dishonestly. A neutral agent (i.e., $\theta_i(t_i) = N$) does not

have an innate behavioral mode. We define $T_i^P \equiv \{t_i \in T_i \mid \theta_i(t_i) = P\}$ and $T^P \equiv T_1^P \times T_2^P$. Similarly, we defined T_i^A, T^A, T_i^N, and T^N.

The partnership has a type profile (t_1, t_2) with the probability of $\pi(t_1, t_2)$. We assume $\pi_i(t_i) \equiv \sum_{t_{i+1} \in T_{i+1}} \pi(t_1, t_2) > 0$ for all $i \in N$ and $t_i \in T_i$.[3] The conditional probability is denoted by $\pi_i(t_{i+1} \mid t_i) = \dfrac{\pi(t_1, t_2)}{\pi_i(t_i)}$.

We have distinct interpretations of π from a socioeconomic perspective and from a strategic perspective. From a socio-economic perspective, we consider π as a model of iterative social interaction in which in any round of social interaction, each type t_i meets each type t_{i+1} in real life $\pi_i(t_{i+1} \mid t_i)$ of the time. From a strategic perspective, we consider π as a model of epistemology, where each type t_i is expected to meet each type t_{i+1} just once with the probability of $\pi_i(t_{i+1} \mid t_i)$.

3. Socio-economic perspective

This section investigates partnerships from a socioeconomic perspective. We assume conformity because a neutral agent is influenced by partners' behavioral modes through social interactions. A neutral agent will behave honestly (dishonestly) if partners are more likely (less likely) to behave honestly than dishonestly. Neither prosocial nor adversarial agents are influenced by partners' behavioral modes.

If partners, even if neutral, are more likely to behave honestly, a neutral agent will also behave honestly. Suppose that an agent is neutral and meets partners who are prosocial by 40%, adversarial by 30%, and neutral by 30%. If the neutral partners remain neutral, the agent will behave honestly because 40% > 30%; that is, partners are more likely to behave honestly on average. If neutral partners behave dishonestly, the agent will behave dishonestly because 40% < 30% + 30%, that is, partners are less likely to behave honestly on average.

[3] We express by $z+1$ the element that is different from z.

To formally describe whether neutral agents follow particular modes and, if so, which modes they follow, we define the following infinite sequence $(\theta_1^k, \theta_2^k)_{k=1}^{\infty}$: for every $i \in \{1,2\}$;

$$\theta_i^k : T_i \to \{P, A, N\} \text{ for all } k \geq 1,$$

$$\theta_i^1 = \theta_i,$$

for every $k \geq 2$ and $t_i \in T_i$,

$$\theta_i^k(t_i) = P \qquad \text{if } \theta_i(t_i) \neq A, \text{ and either } \theta_i(t_i) = P \text{ or } \pi_i(\theta_j^{k-1}(t_j) = P \mid t_i) > \pi_i(\theta_j^{k-1}(t_j) = A \mid t_i),$$

$$\theta_i^k(t_i) = A \qquad \text{if } \theta_i(t_i) \neq P, \text{ and either } \theta_i(t_i) = A \text{ or } \pi_i(\theta_j^{k-1}(t_j) = P \mid t_i) < \pi_i(\theta_j^{k-1}(t_j) = A \mid t_i),$$

and

$$\theta_i^k(t_i) = N \qquad \text{if } \theta_i(t_i) = N, \text{ and } \pi_i(\theta_j^{k-1}(t_j) = P \mid t_i) = \pi_i(\theta_j^{k-1}(t_j) = A \mid t_i).$$

Associated with $(\theta_1^k, \theta_2^k)_{k=1}^{\infty}$, we define an infinite sequence $(T_i^{P,k})_{k=1}^{\infty}$ by

$$T_i^{P,k} \equiv \{t_i \in T_i \mid \theta_i^k(t_i) = P\} \text{ for each } k \geq 1.$$

Similarly, we define $(T_i^{A,k})_{k=1}^{\infty}$ and $(T_i^{N,k})_{k=1}^{\infty}$. For every $k \geq 2$,

$$T_i^{P,k} = T_i^P \bigcup \{t_i \in T_i^N \mid \pi_i(T_{i+1}^{P,k-1} \mid t_i) > \pi_i(T_{i+1}^{A,k-1} \mid t_i)\}\},$$

$$T_i^{A,k} = T_i^A \bigcup \{t_i \in T_i^N \mid \pi_i(T_{i+1}^{P,k-1} \mid t_i) < \pi_i(T_{i+1}^{A,k-1} \mid t_i)\}\},$$

and

$$T_i^{N,k} = \{t_i \in T_i^N \mid \pi_i(T_{i+1}^{P,k-1} \mid t_i) = \pi_i(T_{i+1}^{A,k-1} \mid t_i)\}\}.$$

At the end of the k-th round of iterative social interaction, any neutral agent i who belongs to $T_i^{P,k}$ ($T_i^{A,k}$), that is, meets prosocial (adversarial) partners more often during the (k-1)th round, that is, $\pi_i(T_{i+1}^{P,k-1} \mid t_i) > \pi_i(T_{i+1}^{A,k-1} \mid t_i)$ ($\pi_i(T_{i+1}^{P,k-1} \mid t_i) < \pi_i(T_{i+1}^{A,k-1} \mid t_i)$), follows the prosocial (adversarial) behavioral mode and behaves honestly (dishonestly).

3. 1. Stable society

A society's (T,π,θ) is said to be stable if, for each $i\in\{1,2\}$, there exists $\theta_i^{\infty}:T_i\to\{P,A,N\}$ such that

$$\theta_i^{\infty}(t_i)=\lim_{k\to\infty}\theta_i^{k}(t_i) \text{ for all } t_i\in T_i.$$

Let $K\equiv|T|$. From the finiteness of T, it follows that the society is stable if and only if

$$\theta_i^{K+1}(t_i)=\theta_i^{K}(t_i) \text{ for all } t_i\in T_i,$$

where $\theta_i^{K}(t_i)=\theta_i^{\infty}(t_i)$. The following proposition is self-evident from the definitions.

Proposition 1: A society is stable if and only if for each $i\in N$,

$$T_i^{P,K+1}=T_i^{P,K},\ T_i^{A,K+1}=T_i^{A,K}, \text{ and } T_i^{N,K+1}=T_i^{N,K}.$$

If a society is stable, after iterative social interaction, any neutral agent belonging to $T_i^{P,K}$ follows prosocial agents' behavioral mode, whereas any neutral agent belonging to $T_i^{A,K}$ follows adversarial agents' behavioral mode. If a neutral agent belongs to $T_i^{N,K}$, they do not follow any behavioral mode and remains neutral even after iterative social interaction.

The following example is helpful to understand whether a society is stable or not.

Example 1: Consider a society with symmetry in which

$$T_1=T_2=\{0,1,2,3\},$$

$$\pi(l,l+1)=\pi(l+1,l)>0 \text{ for all } l\in\{0,1,2\},$$

$$[\,|l-l'|\neq 1\,]\Rightarrow[\,\pi(l,l')=0\,] \text{ for all } (l,l')\in\{0,1,2,3\}^2,$$

$$\pi(0,1)\neq\pi(1,2)\neq\pi(2,3),$$

and for each $i\in\{1,2\}$,

$$\theta_i(0)=A,\ \theta_i(3)=P, \text{ and } \theta_i(1)=\theta_i(2)=N.$$

Each type is connected only to its neighbors. All types, except for the polars of the network line, are neutral. The society is not stable if and only if the probability of a neutral partnership (i.e., either $(1,2)$ or $(2,1)$) is the greatest among all possibilities; that is,

$$\pi(1,2) > \max[\pi(0,1), \pi(2,3)].$$

In fact, we have oscillation such that

$$T_i^{P,k} = \{2,3\} \text{ and } T_i^{A,k} = \{0,1\} \text{ for all } k \in \{2,4,6,...\},$$

whereas

$$T_i^{P,k} = \{1,3\} \text{ and } T_i^{A,k} = \{0,2\} \text{ for all } k \in \{3,5,7,...\}.$$

In this case, types 1 and 2 strongly influence each other; if their modes are different, they will switch in the next round, implying this oscillation. Otherwise (i.e., if $\pi(1,2) < \max[\pi(0,1), \pi(2,3)]$), the society is stable. In fact, we have a convergence such that

$$[\pi(1,2) < \min[\pi(0,1), \pi(2,3)]]$$

$$\Rightarrow [T_i^{P,k} = \{2,3\} \text{ and } T_i^{A,k} = \{0,1\} \text{ for all } k \geq 2],$$

$$[\pi(0,1) < \pi(1,2) < \pi(2,3)] \Rightarrow [T_i^{P,k} = \{1,2,3\} \text{ for all } k \geq 3],$$

and

$$[\pi(0,1) > \pi(1,2) > \pi(2,3)] \Rightarrow [T_i^{A,k} = \{1,2,3\} \text{ for all } k \geq 3].$$

If type 1 and type 2 influence each other only weakly, their modes remain unchanged. If either of them unilaterally and strongly influences the other, the mode of the former becomes dominant.

3. 2. Honest society

A society (T, π, θ) is said to be honest if it is stable and, for every $i \in \{1,2\}$ and $t_i \in T_i$,

$$[\theta_i(t_i) = N] \Rightarrow [\theta_i^{\infty}(t_i) = P].$$

We similarly define a dishonest society. The following proposition is self-evident from these definitions.

Proposition 2: A society is honest if and only if

$$T_i^{P,K} = T_i^P \bigcup T_i^N \text{ for each } i \in N.$$

In an honest society, all neutral agents behave honestly after iterative social interaction. In Example 1, a society is honest if and only if $\pi(0,1) < \pi(1,2) < \pi(2,3)$, whereas it is dishonest if and only if $\pi(0,1) > \pi(1,2) > \pi(2,3)$. The following example shows that even if agents are more likely to be adversarial and rarely prosocial, a society can be honest.

Example 2: Assume an arbitrary positive integer $L \geq 1$. Let

$$T_1 = T_2 = \{0,1,...,L,L+1,...,2L\},$$

$$\pi(l,l+1) = \pi(l+1,l) > 0 \text{ for all } l \in \{0,...,2L-1\},$$

$$[\,|l-l'| \neq 1\,] \Rightarrow [\,\pi(l,l') = 0\,] \text{ for all } (l,l') \in \{0,...,2L\}^2,$$

and

$$\pi(0,1) < \pi(1,2) < \cdots < \pi(L-1,L) < \pi(L,L+1)$$
$$> \pi(L+1,L+2) > \cdots > \pi(2L-1,2L).$$

Each type is connected only to its neighbors, and the network line has a single peak at $(L+1,L+2)$. For each $i \in \{1,2\}$, let

$$\theta_i(0) = A,\ \theta_i(2L) = P, \text{ and } \theta_i(l) = N \text{ for all } l \in \{1,...,2L-1\},$$

That is, adversarial and prosocial agents are positioned in the polars of the network line. For each $k \in \{1,...,L-1\}$, we have

$$T_i^{P,k} = \{2L-k,...,2L\} \text{ and } T_i^{A,k} = \{0,1,...,k\}.$$

Since $\pi(L-1,L) < \pi(L,L+1)$, for each $k \in \{L,...,2L-1\}$, we have

$$T_i^{P,k} = \{2L-k,...,2L\} \text{ and } T_i^{A,k} = \{0,1,...,2L-k-1\},$$

and therefore, for each $k \geq 2L-1$, we have

$$T_i^{P,k} = \{1,...,2L\} = T_i^P \bigcup T_i^N.$$

Thus, the society is honest. Because the value of $2\pi(2L-1,2L)$ is not constrained by anything other than being positive, we can see that even if agents are more likely to be adversarial, whereas they are rarely prosocial, a society can be honest. To attain an honest society, it is important that the contagion of the prosocial mode reaches the peak of the network line $(L+1,L+2)$ sooner than that of the adversarial mode.

The following example shows that whether a stable society is honest or dishonest is sensitive to the shape of the distribution π.

Example 3: Let us slightly modify Example 2 by replacing $\pi(0,1)$ and $\pi(0,3)$. Hence, adversarial agent 0 has a connection not with their neighbor (i.e., agent 1) but with a non-neighbor (i.e., agent 3). With this slight change, the contagion of the adversarial mode changes to a peak earlier than that of the prosocial mode. Hence, in contrast to Example 2, we have

$$T_i^{A,1} = \{0,3\},\ T_i^{A,2} = \{0,2,3,4\},\ T_i^{A,3} = \{0,1,2,3,4,5\},$$

$$T_i^{A,k} = \{0,1,...,k+2\} \text{ for all } k \in \{4,...,L-2\},$$

and

$$T_i^{P,k} = \{2L-k,...,2L\} \text{ for all } k \in \{1,...,L-2\}.$$

Since $\pi(L,L+1) > \pi(L+1,L+2)$, for each $k \in \{L-1,...,2L-3\}$, we have

$$T_i^{P,k} = \{k+3,...,2L\} \text{ and } T_i^{A,k} = \{0,1,...,k+2\},$$

that is,

$$T_i^{A,k} = \{0,...,2L-1\} = T_i^P \bigcup T_i^N \text{ for all } k \geq 2L-3.$$

Hence, with a slight change in the linkage of adversarial agent 0, the society becomes dishonest rather than honest.

4. Strategic perspective

This section investigates partnerships from the strategic perspective. We consider π as a model of epistemology, where each type t_i expects their partner to have type t_{i+1} with the probability of $\pi_i(t_{i+1} \mid t_i)$. We assume no social interaction, whereas we assume that a neutral agent behaves strategically according to their selfish motive and epistemological (i.e., rational) reasoning.

Before presenting the details of this section's model, we demonstrate an epistemological condition that is closely related to an honest society. We call a subset of

type profiles $E \subset T$ an event. For convenience, we write $\pi_i(E \mid t_i) \equiv \pi_i(E_{i+1}(t_i) \mid t_i)$ where we denote $E_{i+1}(t_i) \equiv \{t_{i+1} \in T_{i+1} \mid (t_i, t_{i+1}) \in E\}$. Consider an arbitrary event $E \subset T$. Let

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

which denotes the set of agent $i's$ types that know the emergence of E. Let

$$V_i^2(E) = \{t_i \in T_i \mid \pi_i(V_1^1(E) \times V_2^1(E) \mid t_i) = 1\},$$

which denotes the set of agent $i's$ types that know the emergence of $V_1^1(E) \times V_2^1(E)$. For each positive integer, $k \geq 3$, let

$$V_i^k(E) = \{t_i \in T_i \mid \pi_i(V_1^{k-1}(E) \times V_2^{k-1}(E) \mid t_i) = 1\},$$

which denotes the set of agent $i's$ types that know the emergence of $V_1^{k-1}(E) \times V_2^{k-1}(E)$. We define $V_i^\infty(E) \equiv \bigcap_{k=1}^{\infty} V_i^k(E)$. $E \subset T$ is said to be *common knowledge* in a type profile $t \in T$ if $t \in V_1^\infty(E) \times V_2^\infty(E)$. If E is common knowledge at $t \in T$, then $\pi_i\left(V_1^\infty(E) \times V_2^\infty(E) \middle| t_i\right) = 1$ for all $i \in \{1,2\}$.

Proposition 3: If a society is honest, then "no agent is prosocial" never happens to be common knowledge between neutral agents; that is,

(1) $$V_1^\infty(\underset{i \in \{1,2\}}{\times}(T_i^A \cup T_i^N)) \times V_2^\infty(\underset{i \in \{1,2\}}{\times}(T_i^A \cup T_i^N)) \subset T^A.$$

Suppose that no type is adversarial, that is,

(2) $$T_i^A = \phi \text{ for each } i \in \{1,2\}.$$

Then, a society is honest if and only if "both agents are neutral" is never common knowledge; that is,

(3) $$V_1^\infty(T^N) \times V_2^\infty(T^N) = \phi.$$

Proof: Because $V_i^\infty(\underset{j \in \{1,2\}}{\times}(T_j^A \cup T_j^N)) \subset T_i^{A,K} \cup T_i^{N,K}$, it follows that for a society to be honest, it must permit no neutral type to belong to $V_i^\infty(\underset{j \in \{1,2\}}{\times}(T_j^A \cup T_j^N))$, implying (1).

From (2), we have

$$V_i^{\infty}(T^N) = V_i^{\infty}(\underset{j\in\{1,2\}}{\times}(T_j^A \bigcup T_j^N)) \subset T_i^{A,K} \bigcup T_i^{N,K} = T_i^{N,K},$$

which, along with (1), implies that for a society to be honest, $V_i^{\infty}(T^N)$ must be empty; that is, the equality of (3) must hold. The remaining 'if' part is self-evident, because the equality of (3), along with the definition of $T_i^{P,K}$, guarantees the convergence of $(T_i^{P,k})_{k=1}^{\infty}$ on $T_i^P \bigcup T_i^N$. **Q.E.D.**

Matsushima (2022a) terms the condition of epistemology implied by (3) as NCKS. Proposition 3 shows that if prosocial agents exist in epistemology (i.e., NCKS holds) but adversarial agents never exist, then an honest society will be established. However, even if adversarial agents do exist, an honest society can be established, as Examples 2 and 3 indicate. Thus, the honesty of a society is a broader concept than NCKS.

4. 1. Information elicitation

To show the equivalence between the socioeconomic and strategic perspectives, we introduce an information elicitation problem. Let $\Omega = \{0,1\}$ denote a set of states. Each agent knows about the state that occurs, whereas the central planner is unaware of it. The central planner attempts to elicit the correct information from agents by overcoming the fear of their manipulation. We assume that each agent is indifferent to what the central planner uses for information.

From a socioeconomic perspective, if a society is honest, not only prosocial agents but also neutral agents behave honestly and reveal the state truthfully. Hence, without any additional incentive device, the central planner can elicit the correct information (almost certainly) by listening directly to the agents.

From a strategic perspective, by contrast, the central planner does not expect a neutral agent to acquire a prosocial mode. Hence, the central planner must incentivize neutral agents to behave truthfully by designing a (decentralized) mechanism denoted by (M, x), $M = M_1 \times M_2$, $M_i = \underset{k=1}{\overset{K+1}{\times}} M_i^k$, $M_i^k = [0,1]$, $x = (x_i)_{i\in N}$, and $x_i : M \to R$, which denotes the side-payment rule for agent i. The central planner asks each agent the same

question many times (i.e., $K+1$ times) under imperfect information, such as "how likely is state 1 to occur." Each sub-message, $m_i^k \in M_i^k = [0,1]$ and $k \in \{1,2,...,K+1\}$, of agent i expresses the likelihood that state 1 occurs. Each agent is permitted to announce different opinions about this likelihood across the sub-messages. Depending on their announcements $m = (m_1, m_2) \in M$, the central planner makes a side-payment to agent i, which is given by $x_i(m) \in R$.

A strategy for agent i is defined as $s_i = (s_i^k)_{k=1}^{K+1}$, where $s_i^k : \Omega \times T_i \to M_i^k$ for each $k \in \{1,2,...,K+1\}$. At each state $\omega \in \{0,1\}$, agent i with type t_i announces $m_i^k = s_i^k(\omega, t_i) \in [0,1]$ as their k-th sub-message. We assume that a prosocial agent reveals the state honestly, while an adversarial agent reveals it dishonestly. Hence, we confine our attention to strategies s_i such that for every $\omega \in \{0,1\}$ and $k \geq 1$,

$$s_i^k(\omega, t_i) = \omega \text{ for all } t_i \in T_i^P,$$

and

$$s_i^k(\omega, t_i) = \omega + 1 \text{ for all } t_i \in T_i^A.$$

A neutral agent i (i.e., $\theta_i(t_i) = N$) is selfish and therefore maximizes the expected value of side-payment $x_i(m)$. A strategy profile s is said to be a BNE in the associated game with society (T, π, θ) and mechanism (M, x) if for every $i \in N$, $t_i \in T_i^N$ (i.e., $\theta_i(t_i) = N$), $\omega \in \Omega$, and $m_i \in M_i$,

$$E[x_i(s_i(\omega, t_i), m_{i+1}) \mid \omega, s_{i+1}, t_i] \geq E[x_i(m_i, m_{i+1}) \mid \omega, s_{i+1}, t_i].^4$$

4. 2. Solvable mechanism and implementation

A mechanism (M, x) is said to be solvable in a society (T, π, θ) if there exists a unique BNE s, and this unique BNE s satisfies that

$$s_i^{K+1} = s_i^K \text{ for all } i \in \{1,2\}.$$

[4] $E[\cdot \mid \omega, s_{i+1}, t_i]$ expresses the expectation operator conditional on (ω, s_{i+1}, t_i).

In a solvable society, all types who play the unique BNE eventually stop changing their announcements.

We define $I:[0,1]\to[0,1]$ as

$$I(\alpha)=1 \qquad \text{if } \alpha>1/2,$$

$$I(\alpha)=1/2 \qquad \text{if } \alpha=1/2,$$

and

$$I(\alpha)=0 \qquad \text{if } \alpha<1/2.$$

The central planner interprets each agent $i's$ k-th sub-message m_i^k ($\neq 1/2$) as the occurrence of the state $\omega=I(m_i^k)\in\{0,1\}$. If $I(m_i^k)=1/2$, the central planner considers m_i^k to be not informative about the state.

The central planner takes each agent's final announcement (i.e., m_i^{K+1}) as their true intention. A mechanism (M,x) is said to implement a society (T,π,θ) if it is solvable, and the associated unique BNE s satisfies that for every $i\in N$, $\omega\in\{0,1\}$, and $t_i\in T_i$,

$$[\,I(s_i^{K+1}(\omega,t_i))=\omega\,]\Leftrightarrow[\,\theta_i^K(t_i)=P\,],$$

$$[\,I(s_i^{K+1}(\omega,t_i))=\omega+1\,]\Leftrightarrow[\,\theta_i^K(t_i)=A\,],$$

and

$$[\,I(s_i^{K+1}(\omega,t_i))=1/2\,]\Leftrightarrow[\,\theta_i^K(t_i)=N\,].$$

If the mechanism implements a society, any neutral agent's final announcement (i.e., $m_i^{K+1}=s_i^{K+1}(\omega,t_i)$) in the unique BNE is consistent with the behavioral mode acquired through iterative social interaction.

4. 3. Equivalence

For each $i\in\{1,2\}$, we specify the side-payment rule x_i as the following combination of the quadratic scoring rule (i.e., $(m_i^1-1/2)^2$) and its variant (i.e., $\{m_i^k-I(m_{i+1}^{k-1})\}^2$):

$$x_i(m) = \varepsilon(m_i^1 - 1/2)^2 + \sum_{k=2}^{K+1} \varepsilon^k \{m_i^k - I(m_{i+1}^{k-1})\}^2 \quad \text{for all } m \in M .$$

The specified mechanism (M, x) is detail-free, that is, it does not depend on the details of the specification of the society. We show that the stability of a society is equivalent to its implementation using the specified mechanism.

Theorem 1: For every society (T, π, θ), the following three properties are equivalent:

(i) The society (T, π, θ) is stable.

(ii) The mechanism (M, x) is solvable in the society (T, π, θ).

(iii) The mechanism (M, x) implements the society (T, π, θ).

Proof: A neutral agent i selects m_i^1 that maximizes $\varepsilon(m_i^1 - 1/2)^2$, that is, they select $m_i^1 = 1/2$ uniquely, because it is the only part of the side-payment rule that is relevant to their first sub-message. Hence, we have

$$s_i^1(\omega, t_i) = 1/2 \text{ and } I(s_i^1(\omega, t_i)) = 1/2 \quad \text{if agent } i \text{ is neutral.}$$

We also have

$$[\, I(s_i^1(\omega, t_i)) = 1/2 \,] \Leftrightarrow [\, \theta_i^1(t_i) = N \,]$$

and

$$[\, I(s_i^1(\omega, t_i)) = \omega \,] \Leftrightarrow [\, \theta_i^1(t_i) = P \,].$$

For each $k \geq 2$, a neutral agent i selects m_i^k that maximizes the expected value of $\varepsilon^k \{m_i^k - I(m_{i+1}^{k-1})\}^2$, that is, they select $m_i^k = E[I(s_{i+1}^{k-1}(\omega, t_{i+1})) \mid \omega, t_i]$ uniquely, because it is the only part of the side-payment rule that is relevant to the k-th sub-message. Hence, we have

$$s_i^k(\omega, t_i) = E[I(s_{i+1}^{k-1}(\omega, t_{i+1})) \mid \omega, t_i] \quad \text{if agent } i \text{ is neutral.}$$

If

$$[\, I(s_{i+1}^{k-1}(\omega, t_{i+1})) = 1/2 \,] \Leftrightarrow [\, \theta_{i+1}^{k-1}(t_{i+1}) = N \,],$$

$$[\, I(s_{i+1}^{k-1}(\omega, t_{i+1})) = \omega \,] \Leftrightarrow [\, \theta_{i+1}^{k-1}(t_{i+1}) = P \,],$$

and

$$[\, I(s_{i+1}^{k-1}(\omega, t_{i+1})) = \omega + 1\,] \Leftrightarrow [\, \theta_{i+1}^{k-1}(t_{i+1}) = A\,],$$

then, we have

$$[\, I(s_i^k(\omega, t_i)) = 1/2\,] \Leftrightarrow [\, \theta_i^k(t_i) = N\,],$$

$$[\, I(s_i^k(\omega, t_i)) = \omega\,] \Leftrightarrow [\, \theta_i^k(t_i) = P\,],$$

and

$$[\, I(s_i^k(\omega, t_i)) = \omega + 1\,] \Leftrightarrow [\, \theta_i^k(t_i) = A\,].$$

From these observations, we have proven that the specified mechanism is solvable in a society if and only if this society is stable. In this case, the associated unique BNE satisfies

$$[\, I(s_i^{K+1}(\omega, t_i)) = 1/2\,] \Leftrightarrow [\, \theta_i^K(t_i) = N\,]$$

$$[\, I(s_i^{K+1}(\omega, t_i)) = \omega\,] \Leftrightarrow [\, \theta_i^K(t_i) = P\,]$$

and

$$[\, I(s_i^{K+1}(\omega, t_i)) = \omega + 1\,] \Leftrightarrow [\, \theta_i^K(t_i) = A\,].$$

That is, if a society is stable, the specified mechanism (M, x) of automatically implements it. Hence, we have proven the equivalence of (i), (ii), and (iii).

Q.E.D.

Corollary 1: A society is honest if and only if the mechanism (M, x) implements it and the associated unique BNE s satisfies

$$s_i^{K+1}(\omega, t_i) = \omega \quad \text{for all} \quad \omega \in \{0,1\} \quad \text{and} \quad t_i \in T_i^N.$$

Theorem 1 and Corollary 1 imply that the central planner can solve the information elicitation problem in a society from the strategic perspective if and only if the society is honest. We must be aware that even when considering different detail-free mechanisms, the central planner cannot expand the range of societies in which neutral agents prefer to behave honestly beyond the set of honest societies without harm. Suppose that an arbitrary detail-free mechanism is solvable and that it induces neutral agents to behave honestly in a society that is not honest. Consider another society specified by swapping prosocial types with adversarial types. This society is not dishonest. However, owing to

the detail-freeness, neutral agents turn into liars in the mechanism. Thus, the range of societies in which neutral agents prefer to behave dishonestly is also expanded beyond the set of dishonest societies at the same time.

Matsushima (2022a) considers the case in which each agent is either prosocial or neutral and shows a positive result in that the central planner can solve the information elicitation problem from the strategic perspective if NCKS holds. Theorem 1 and Corollary 1 generalize this result by considering adversarial agents in addition to prosocial agents and neutral agents.

In the presence of adversarial types, we need a different mechanism design from that of Matsushima (2022a). In Matsushima (2022a), owing to the non-existence of adversarial types, it suffices for the central planner to ask an agent for a single answer. However, to eliminate the strategic impacts of adversarial agents on neutral agents' epistemological consideration, we introduce a new mechanism design method according to which the central planner asks the same questions to each agent repeatedly and adopts the final answers as their true intention. By mounting a nested structure of the quadratic scoring rule and its variants, this method can gradually eliminate the influence of adversarial agents and induce neutral agents to reveal the truth after multiple announcements.

The mechanism design in this study is similar to that in Matsushima (2022b) in that it incorporates multiple announcements and a nested structure; however, the two are fundamentally different. Matsushima (2022b) generalizes the complete information studied by Matsushima (2022a) to asymmetric information. Unlike complete information, no one else knows the same thing, which dissuades each agent from being honest on the first attempt. By inducing agents to speak frequently through the nested structure of the quadratic scoring rule and its variant, the central planner succeeds in gradually expanding the number of selfish types that have an incentive to be honest, which solves the implementation problem of social choice functions.

5. Conclusion

In this study, we investigate a simple form of information elicitation problem from socioeconomic and strategic perspectives. We demonstrate equivalence between these

perspectives by showing that the central planner can uniquely elicit the correct information in a detail-free mechanism if and only if a society is honest; that is, the majority of agents can acquire the prosocial behavioral mode after iterative social interaction with conformity. We permit adversarial agents to exist in social networks, which makes the solution to the problem theoretically essential. We then demonstrate a new mechanism design method for solving this problem.

We simplified the problem to make it easier to understand the logical essence of this study. However, careful consideration is necessary to determine the extent to which the conclusions can be generalized. For example, three or more-agent partnerships, three or more possible states, and asymmetric information on the state can be considered. Moreover, a general implementation problem of social choice functions instead of the specific information elicitation problem can be investigated, by removing the assumption that each agent is indifferent to how the central planner uses the collected information. Although not analyzed explicitly, we hope that we can extend the consequences of this study to these cases without substantial difficulty. In fact, Matsushima (2022a, 2022b) considers the case in which each agent is either neutral or honest but not adversarial, and provides positive results for the general implementation of social choice functions. The solution to the information elicitation problem is crucial in solving all implementation problems.

Matsushima (2022a) considers weakly honest agents to weigh honesty against self-interest. Considering weakly adversarial agents similarly, a new argument must be created in future research. Adversarial agents are more likely to behave honestly when they are surrounded by a large number of (innate or acquired) prosocial agents. We should also consider the possibility that a type includes information about the differences in how social influences are received from and given to others.

Another future research direction would be to clarify the social impacts of individual agents and the social planner's activities on shaping the social network. This research agenda is related to the literature on network formation (Bala and Goyal, 2000) and social mobility (Chetty et al., 2022a, 2022b). Example 3 shows that a slight change in an adversarial agent's ties in a social network has a significant impact on well-being and performance. It would be interesting to investigate what kinds of social impacts arise in terms of forming social practices, social norms, and social common capital as well as

promoting human capital and economic success. The above-mentioned future research themes demonstrate the limitations of this study and require the formation of a more careful model in which the central planner is more aware of the details of the specification of a society.

Acknowledgements

This study was supported by the Japan Society for the Promotion of Science (KAKENHI 20H00070); the Japanese Ministry of Education, Culture, Sports, Science, and Technology; and the Endowed Chair on Social Common Capital (Ryohin Keikaku Co. Ltd.) at the Department of Economics, University of Tokyo. The authors do not have any conflict of interest.

References

Abeler, J., Nosenzo, D., Raymond, C., 2019. Preference for truth-telling. Econometrica. 87, 1115–1153.

Bala, V., Goyal, S., 2000. A noncooperative model of network formation. Econometrica. 68, 1181–1229.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3.

Burt, R., 2018. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, Massachusetts.

Chetty, R., Jackson, M.O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R.B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., Barberá, P., Bhole, M., Wernerfelt, N., 2022a. Social capital I: Measurement and associations with economic mobility. Nature. 608, 108–121.

Chetty, R., Jackson, M.O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R.B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R.,

Bailey, M., Barberá, P., Bhole, M., Wernerfelt, N., 2022b. Social capital II: Determinants of economic connectedness. Nature. 608, 122–134.

Christakis, N.A., J. Fowler, 2007. The Spread of Obesity in a Large Social Network over 32 Years. N. Engl. J. Med. 357, 370–379.

Coleman, J.S., 1988. Social capital in the creation of human capital. Am. J. Sociol. 94, S95–S120.

Dutta, B., Sen, A., 2012. Nash implementation with partially honest individuals. Games Econ. Behav. 74, 154–169.

Granovetter, M.S., 1973. The Strength of Weak Ties. Am. J. Sociol. 78, 1360–1380.

Kartik, N., Tercieux, O., Holden, R., 2014. Simple mechanisms and preferences for honesty. Games Econ. Behav. 83, 284–290.

Kohler, H.P., J. Behrman, Watkins, S.C., 2001. The Density of Social Networks and Fertility Decisions: Evidence from South Nyanza District, Kenya. Demography. 38, 43–58.

Matsushima, H., 2008a. Behavioral aspects of implementation theory. Econ. Lett. 100, 161–164.

Matsushima, H., 2008b. Role of honesty in full implementation. J. Econ. Theor. 139, 353–359.

Matsushima, H., 2013. Process manipulation in unique implementation. Soc. Choice Welf. 41, 883–893.

Matsushima, H., 2022a. Epistemological implementation of social choice functions. Games Econ. Behav. 136, 389-402.

Matsushima, H., 2022b. Honesty and epistemological implementation with asymmetric information. SSRN Journal. http://dx.doi.org/10.2139/ssrn.4177182.

Mukherjee, S., Muto, N., Ramaekers, E., 2017. Implementation in undominated strategies with partially honest agents. Games Econ. Behav. 104, 613–631.

Ortner, J., 2015. Direct implementation with minimally honest individuals. Games Econ. Behav. 90, 1–16.

Putnam, R.D., 2000. Bowling Alone. The Collapse and Revival of American Community, New York. Touchstone.

Putnam, R., 2004. Democracies in Flux: The Evolution of Social Capital in Contemporary Society. Oxford University Press, Oxford.

Rosenquist, J.N., J. Fowler, N. Christakis, 2011. Social network determinants of depression. Mol. Psychiatry. 16, 273–281.

Saporiti, A., 2014. Securely implementable social choice rules with partially honest agents. J. Econ. Theor. 154, 216–228.

Uzawa, H., 2005, Economic Analysis of Social Common Capital. Cambridge University Press, Cambridge, England.