The revelation principle does not always hold when strategies of agents are costly

Wu, Haoyang

16 September 2018

The revelation principle does not always hold when strategies of agents are costly

Haoyang Wu *

Wan-Dou-Miao Research Lab, Room 301, Building 3, 718 WuYi Road, 200051, China.

Abstract

The revelation principle asserts that for any indirect mechanism and equilibrium, there is a corresponding direct mechanism with truth as an equilibrium. Although the revelation principle has been a fundamental theorem in the theory of mechanism design for a long time, so far the costs related to strategic actions of agents have not been fully discussed. In this paper, we propose the notion of profit function, and claim that the definition of Bayesian Nash equilibrium of mechanism should be based on the profit function instead of the utility function when strategies of agents are costly (see Definition 23.D.1'). After then, we derive two key results: (1) The strategic action of each agent in a direct mechanism is just to report a type, and each agent does not need to spend any strategic cost occurred in any indirect mechanism (see Proposition 1); (2) When strategies of agents are costly, the proof of revelation principle is wrong (see Proposition 2). We construct a simple labor model to show that a Bayesian implementable social choice function is not truthfully implementable (see Proposition 4), which contradicts the revelation principle.

JEL codes: D71, D82

Key words: Revelation principle; Game theory; Mechanism design.

1 Introduction

The revelation principle is a fundamental theorem in mechanism design theory [1–3]. According to the wide-spread textbook given by Mas-Colell, Whinston and Green (Page 884, Line 24 [3]): "*The implication of the revelation principle is ... to identify the set of implementable social choice functions in*

* Corresponding author.
Email address: 18621753457@163.com (Haoyang Wu).

Bayesian Nash equilibrium, we need only identify those that are truthfully implementable." Put in other words, the revelation principle says: "*suppose that there exists a mechanism that implements a social choice function f in Bayesian Nash equilibrium, then f is truthfully implementable in Bayesian Nash equilibrium*" (Page 76, Theorem 2.4, [4]). Relevant definitions about the revelation principle are given in Section 2, which are cited from Section 23.B and 23.D of MWG's textbook [3].

Generally speaking, agents may spend some costs when participating a mechanism. There are two kinds of costs possibly occurred in a mechanism: 1) *strategic costs*, which are possibly spent by agents when performing strategic actions [1]; 2) *misreporting costs*, which are possibly spent by agents when reporting types falsely. [2] In the traditional literature of mechanism design, costs are usually referred to the former. Recently, some researchers began to investigate misreporting costs. For every type θ and every type $\hat{\theta}$ that an agent might misreport, Kephart and Conitzer [6] defined a cost function as $c(\theta, \hat{\theta})$ for doing so. Traditional mechanism design is just the case where $c(\theta, \hat{\theta}) = 0$ everywhere, and partial verification is a special case where $c(\theta, \hat{\theta}) \in \{0, \infty\}$ [7,8]. Kephart and Conitzer [6] proposed that when reporting truthfully is costless and misreporting is costly, the revelation principle can fail to hold.

Despite these accomplishments, so far people seldom consider the two kinds of costs simultaneously. The aim of this paper is to investigate whether the revelation principle holds or not when two kinds of costs are considered. The paper is organized as follows. In Section 2, we propose the notion of profit function, and claim that the definition of Bayesian Nash equilibrium of mechanism should be based on the profit function instead of the utility function when strategies of agents are costly (see Definition 23.D.1'). After then, we point out two key points:
(1) Each agent's strategy in a direct mechanism is just to report a type. Hence each agent does not need to spend any strategic cost occurred in any indirect mechanism (see Proposition 1);
(2) When strategies of agents are costly, the proof of revelation principle in Proposition 23.D.1 [3] is wrong (see Proposition 2).

In Section 3, we construct a simple labor model, then define a social choice function f and an indirect mechanism, in which strategies of agents are costly. In Section 4, we prove f can be implemented by the indirect mechanism in Bayesian Nash equilibrium. In Section 5, we show that f is not truthfully implementable in Bayesian Nash equilibrium under some conditions (see Proposition 4), which contradicts the revelation principle. In the end, Section 6 draws conclusions.

[1] For example, agents spend education costs in a job market [5].
[2] It is usually assumed that each agent can report his true type with zero cost.

2 Analysis of strategic costs

In this section, we will investigate costs spent by agents when playing strategies in a mechanism. In the beginning ,we cite some definitions from Section 23.B and Section 23.D of MWG's textbook [3] as follows. Consider a setting with I agents, indexed by $i = 1, \cdots, I$. Each agent i privately observes his type θ_i that determines his preferences. The set of possible types of agent i is denoted as Θ_i. The agent i's utility function over the outcomes in set X given his type θ_i is $u_i(x, \theta_i)$, where $x \in X$.

Note 1: Generally speaking, when an agent performs a strategic action in participating a game, he usually spends some monetary costs (or make some efforts which can be quantified as monetary costs). Assume each agent's costs are only relevant to his strategic action and private type, and are independent of the game outcome.

Formally, suppose an agent i with private type $\theta_i \in \Theta_i$ performs a strategic action $s_i(\theta_i)$ and the game outcome is x, then his strategic costs can be denoted as $c_i(s_i(\theta_i), \theta_i)$. Thus, agent i's profit is denoted as

$$p_i(x, s_i(\theta_i), \theta_i) = u_i(x, \theta_i) - c_i(s_i(\theta_i), \theta_i). \tag{1}$$

Definition 23.B.1 [3]: A *social choice function* (SCF) is a function $f : \Theta_1 \times \cdots \times \Theta_I \to X$ that, for each possible profile of the agents' types $\theta_1, \cdots, \theta_I$, assigns a collective choice $f(\theta_1, \cdots, \theta_I) \in X$.

Definition 23.B.3 [3]: A *mechanism* $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ is a collection of I strategy sets $S_1, \cdots, S_I$ and an outcome function $g : S_1 \times \cdots \times S_I \to X$.

Definition 23.B.5 [3]: A *direct mechanism* is a mechanism $\Gamma' = (S'_1, \cdots, S'_I, g'(\cdot))$ in which $S'_i = \Theta_i$ for all i and $g'(\theta) = f(\theta)$ for all $\theta \in \Theta_1 \times \cdots \times \Theta_I$.

Note 2: In a direct mechanism, each agent's report can be considered as an oral and costless announcement: *i.e.*, the strategy of each agent i with private type θ_i is to report a type $s'_i(\theta_i) \in \Theta_i$, and $s'_i(\theta_i)$ needn't to be his private type θ_i. After the designer receives all reports $s'_1(\theta_1), \cdots, s'_I(\theta_I)$, he must announce the outcome $f(s'_1(\theta_1), \cdots, s'_I(\theta_I))$.

Definition 23.D.1 [3]: The strategy profile $s^*(\cdot) = (s^*_1(\cdot), \cdots, s^*_I(\cdot))$ is a *Bayesian Nash equilibrium* of mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ if, for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s^*_i(\theta_i), s^*_{-i}(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s^*_{-i}(\theta_{-i})), \theta_i)|\theta_i] \tag{2}$$

for all $\hat{s}_i \in S_i$.

Note 3: In Definition 23.D.1, the utility function $u_i(x, \theta_i)$ is used to define the Bayesian Nash equilibrium. Usually, in an indirect mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$, each agent i's strategy $s_i(\theta_i)$ is an action that requires some costs to be performed, *i.e.*, $c_i(s_i(\theta_i), \theta_i) > 0$. In this case, the utility function $u_i(x, \theta_i)$ only describes the utility of agent i after obtaining the outcome x but misses his costs, thus cannot describe the whole profit of agent i [3]. Actually, the profit function $p_i(x, s_i(\theta_i), \theta_i)$ should be used to define the Bayesian Nash equilibrium of a mechanism. Put in other words, Definition 23.D.1 should be reformulated as follows:

Definition 23.D.1' The strategy profile $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$ is a *Bayesian Nash equilibrium* of mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ if, for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[p_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), s_i^*(\theta_i), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[p_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \hat{s}_i, \theta_i)|\theta_i] \tag{3}$$

i.e.,

$$\begin{aligned} &E_{\theta_{-i}}[(u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i) - c_i(s_i^*(\theta_i), \theta_i))|\theta_i] \geq \\ &\quad E_{\theta_{-i}}[(u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i) - c_i(\hat{s}_i, \theta_i))|\theta_i] \end{aligned}$$

for all $\hat{s}_i \in S_i$, in which p_i is the profit of agent i given by Eq (1).

Definition 23.D.2 [3]: The mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ *implements the social choice function* $f(\cdot)$ *in Bayesian Nash equilibrium* if there is a Bayesian Nash equilibrium of Γ, $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$, such that $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

Definition 23.D.3 [3]: The social choice function $f(\cdot)$ is *truthfully implementable in Bayesian Nash equilibrium* (or *Bayesian incentive compatible*) if $s_i'^*(\theta_i) = \theta_i$ for all $\theta_i \in \Theta_i$ and $i = 1, \cdots, I$ is a Bayesian Nash equilibrium of the direct revelation mechanism $\Gamma' = (S_1', \cdots, S_I', g'(\cdot))$, in which $S_i' = \Theta_i$, $g' = f$. That is, if for all $i = 1, \cdots, I$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i], \qquad (23.D.1)$$

for all $\hat{\theta}_i \in \Theta_i$.

Note 4: Following Footnote 2, if agent i with private type θ_i chooses $s_i'^*(\theta_i) = \theta_i$ in the direct mechanism $\Gamma' = (S_1', \cdots, S_I', g'(\cdot))$, then $c_i(s_i'(\theta_i), \theta_i) = 0$, and $p_i(x, s_i(\theta_i), \theta_i) = u_i(x, \theta_i)$ by Eq (1). Therefore, when strategic actions of agents are costly, although the notion of Bayesian Nash equilibrium should be

[3] In many practical cases, strategies of agents are costly actions. Only in some restricted cases where strategies of agents are oral announcements can strategies be viewed costless. Thus, the traditional definition of Bayesian Nash equilibrium holds only in these restricted cases.

reformulated from Definition 23.D.1 to Definition 23.D.1' according to Note 3, the notion of Bayesian incentive compatibility can still be defined by Definition 23.D.3.

Proposition 1: The strategic action of each agent i in the direct mechanism $\Gamma' = (S'_1, \cdots, S'_I, g'(\cdot))$ is just to report a type from Θ_i. Each agent i does not need to take any other action to prove himself that his reported type is truthful, and should not play any strategic action as specified in any indirect mechanism. Hence, *in a direct mechanism, each agent does not need to spend strategic costs related to strategic actions specified in any indirect mechanism.*

Proof: As pointed out in Definition 23.B.5, in the direct mechanism Γ', the strategy set $S'_i = \Theta_i$, which means that the strategy s'_i of agent i with private type θ_i is just to choose a type from Θ_i to report, *i.e.*, $s'_i(\theta_i) \in \Theta_i$.

Obviously, the designer cannot enforce each agent to report truthfully in the direct mechanism, and *each agent does not need to take any action to prove himself that his reported type is truthful.* Otherwise, assume to the contrary that each agent i has to submit some additional evidences to the designer in order to prove himself that his reported type is truthful, *i.e.*, $s'_i(\theta_i) = \theta_i$. Then *there will be no information disadvantage from the viewpoint of the designer*: the agents' types $\theta_1, \cdots, \theta_I$ are no longer their private information, and the designer can directly specify his favorite outcome $f(\theta_1, \cdots, \theta_I)$ without any uncertainty after receiving agents' reports. This case contradicts the basic framework of mechanism design, therefore the assumption does not hold.

Hence, each agent i with private type θ_i will misreport another type $s'_i(\theta_i) \neq \theta_i$, $s'_i(\theta_i) \in \Theta_i$ whenever doing so is worthwhile. After the designer receives $s'_1(\theta_1), \cdots, s'_I(\theta_I)$, he has no way to verify whether these reports are truthful or not. What the designer can do is just to announce $f(s'_1(\theta_1), \cdots, s'_I(\theta_I))$ as outcome. Thus, *it is illegal to assume that in a direct mechanism the designer can require each agent perform any strategic action specified in any indirect mechanism.* As a result, in a direct mechanism, each agent i does not need to spend strategic costs related to strategic actions specified in any indirect mechanism. □

Discussion 1: Someone may disagree with Proposition 1 and argue that the notion of direct mechanism can be extended as follows: For a given social choice function f, the designer defines an "extended direct mechanism", in which each agent reports a type, then the designer suggests each agent which action to take, and the final outcome function depends on agents' actions and is just equal to f. As a result, each agent still spend strategic costs, the same as what they spend in the indirect mechanism.

Answer 1: It should be noted that behind the so-called extended direct mechanism, there actually exists an underlying assumption: *Each agent is willing*

to inform full details of his private strategy chosen in the indirect mechanism to the designer. Only when this assumption holds can the designer suggest each agent to take which strategic action after receiving an arbitrary profile of agents' reported types.

However, in the framework of mechanism design, the strategy of each agent i in an indirect mechanism is his private function $s_i : \Theta_i \to S_i$ describing agent i's choice for each possible type in Θ_i that he might have [3]. *The strategy of each agent in an indirect mechanism is his private choice*. It is unreasonable to simply assume that each agent will voluntarily inform full details of his private strategy to the designer without obtaining any more profits. [4] Consequently, the so-called extended direct mechanism does not hold. □

Now we will investigate whether the revelation principle still holds or not when strategies of agents are costly. The revelation principle for Bayesian Nash equilibrium is cited in Appendix. Traditional explanation is as follows: "*Consider the equilibrium in an indirect mechanism, there is a mapping from vectors of agents' types into outcomes. Now suppose we take that mapping to be a revelation game, then no type of any agent can make an announcement that differs from his true type and do better*". We will refute this explanation by the following Proposition 2.

Proposition 2: Given an indirect mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$, if each strategic action $s_i(\theta_i)$ is costly, *i.e.*, $c_i(s_i(\theta_i), \theta_i) > 0$, then the proof of the revelation principle given in Proposition 23.D.1 is wrong.

Proof: According to the proof of Proposition 23.D.1 (see Appendix), suppose that there exists an indirect mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ that implements the social choice function $f(\cdot)$ in Bayesian Nash equilibrium, then there exists a profile of strategies $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$ such that the mapping $g(s^*(\cdot)) : \Theta_1 \times \cdots \times \Theta_I \to X$ from a vector of agents' types $\theta = (\theta_1, \cdots, \theta_I)$ into an outcome $g(s^*(\theta))$ is equal to the desired outcome $f(\theta)$, *i.e.*, $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta_1 \times \cdots \times \Theta_I$. By Definition 23.D.1', for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[(u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i) - c_i(s_i^*(\theta_i), \theta_i))|\theta_i] \geq E_{\theta_{-i}}[(u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i) - c_i(\hat{s}_i, \theta_i))|\theta_i]$$

for all $\hat{s}_i \in S_i$.

Thus, for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[(u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i) - c_i(s_i^*(\theta_i), \theta_i))|\theta_i] \geq E_{\theta_{-i}}[(u_i(g(s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})), \theta_i) - c_i(s_i^*(\hat{\theta}_i), \theta_i))|\theta_i]$$

[4] The notion of direct mechanism defined in MWG's book does not need the so-called assumption (see Definition 23.B.5, [3])

for all $\hat{\theta}_i \in \Theta_i$.

Since $g(s^*(\theta)) = f(\theta)$ for all θ, then for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[(u_i(f(\theta_i, \theta_{-i}), \theta_i) - c_i(s_i^*(\theta_i), \theta_i))|\theta_i] \geq E_{\theta_{-i}}[(u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i) - c_i(s_i^*(\hat{\theta}_i), \theta_i))|\theta_i], \tag{4}$$

for all $\hat{\theta}_i \in \Theta_i$. Note that this inequality *cannot* infer the inequality in Definition 23.D.3, which represents the sufficient condition of Bayesian incentive compatibility. Consequently, the proof of the revelation principle given in Proposition 23.D.1 is wrong. □

3 A labor model and a social choice function f

Here we construct a labor model which uses ideas from the first-price sealed auction model in Example 23.B.5 [3] and the signaling model [3,5]. There are one firm and two agents. Agent 1 and Agent 2 differ in the number of units of output that they produce if hired by the firm, which is denoted by private productivity type. The firm chooses wage $w > 0$ and wants to hire an agent with productivity as high as possible, and the two agents compete for this job.

For simplicity, we make the following assumptions:
1) The possible productivity types of two agents are: θ_L and θ_H, where $\theta_H > \theta_L > 0$. Each agent i's productivity type θ_i $(i = 1, 2)$ is his private information.
2) There is a certificate that the firm can announce as a hire criterion. If each of (or neither of) two agents has the certificate, then each agent will be hired with probability 0.5. The education level corresponding to the certificate is $e_H > 0$. Each agent decides by himself whether to get the certificate or not, hence the possible education level e_i of each agent $i = 1, 2$ is e_H or 0. The education level does nothing for an agent's productivity.
3) The strategic cost of obtaining education level e_i for agent i $(i = 1, 2)$ with productivity type θ_i is given by a function $c_i(e_i, \theta_i) = e_i/\theta_i$. That is, the strategic cost is lower for a higher productivity agent.
4) The misreporting cost for a low-productivity agent to report the high-productivity type θ_H is a fixed value $c_{mis} \geq 0$. In addition, a high-productivity agent is assumed to report the low-productivity type θ_L with zero costs.

The labor model's outcome is represented by a vector (y_1, y_2), where y_i denotes the probability that agent i gets the job. Recall that the firm does not know the exact productivity types of two agents, and its aim is to hire an agent with productivity as high as possible. This aim can be represented by a social choice function $f(\theta) = (y_1(\theta), y_2(\theta))$, in which $\theta = (\theta_1, \theta_2)$, y_i $(i = 1, 2)$ is the

probability that agent i gets the job.

$$y_1(\theta) = \begin{cases} 1, & \text{if } \theta_1 > \theta_2 \\ 0.5, & \text{if } \theta_1 = \theta_2 \\ 0, & \text{if } \theta_1 < \theta_2 \end{cases}, \; y_2(\theta) = \begin{cases} 1, & \text{if } \theta_1 < \theta_2 \\ 0.5, & \text{if } \theta_1 = \theta_2 \\ 0, & \text{if } \theta_1 > \theta_2 \end{cases},$$

$$f(\theta) = (y_1(\theta), y_2(\theta)) = \begin{cases} (1,0), & \text{if } \theta_1 > \theta_2 \\ (0.5,0.5), & \text{if } \theta_1 = \theta_2 \\ (0,1), & \text{if } \theta_1 < \theta_2 \end{cases}. \tag{5}$$

In order to implement the above social choice function $f(\theta)$, the firm designs an indirect mechanism $\Gamma = (S_1, S_2, g)$ as follows: Each agent $i = 1, 2$, conditional on his type $\theta_i \in \{\theta_L, \theta_H\}$, chooses his education level as a bid $e_i : \{\theta_L, \theta_H\} \to \{0, e_H\}$. The strategy set S_i is the set of agent i's all possible bids, and the outcome function g is defined as:

$$g(e_1, e_2) = (g_1, g_2) = \begin{cases} (1,0), & \text{if } e_1 = e_H, e_2 = 0 \\ (0.5,0.5), & \text{if } e_1 = e_2 \\ (0,1), & \text{if } e_1 = 0, e_2 = e_H \end{cases}, \tag{6}$$

where g_i $(i = 1, 2)$ is the probability that agent i gets the job.

Let u_0 be the expected utility of the firm, then $u_0(e_1, e_2) = g_1\theta_1 + g_2\theta_2 - w$. Let u_1, u_2 be the utilities of agent $1, 2$, and p_1, p_2 be the profits of agent $1, 2$ in the indirect mechanism Γ respectively, then for $i, j = 1, 2$, $i \neq j$,

$$u_i(e_i, e_j; \theta_i) = \begin{cases} w, & \text{if } e_i > e_j \\ 0.5w, & \text{if } e_i = e_j \\ 0, & \text{if } e_i < e_j \end{cases}, \tag{7}$$

$$p_i(e_i, e_j; \theta_i) = u_i(e_i, e_j; \theta_i) - c_i(e_i, \theta_i) = u_i(e_i, e_j; \theta_i) - e_i/\theta_i. \tag{8}$$

The item "e_i/θ_i" in Eq (8) stands for the strategic costs spent by agent i with type θ_i when he performs the strategy e_i in the indirect mechanism. [5] Suppose the reserved utilities of agent 1 and agent 2 are both zero, then the individual rationality (IR) constraints are: $p_i(e_i, e_j; \theta_i) \geq 0$, $i = 1, 2$.

4 f is Bayesian implementable

Proposition 3: If $w \in (2e_H/\theta_H, 2e_H/\theta_L)$, the social choice function $f(\theta)$ given in Eq (5) is Bayesian implementable, *i.e.*, it can be implemented by the

[5] For the case of $e_i < e_j$, there will be $e_i = 0$.

indirect mechanism Γ given by Eq (6) in Bayesian Nash equilibrium.
Proof: Consider a separating strategy, *i.e.*, agents with different productivity types choose different education levels,

$$e_1(\theta_1) = \begin{cases} e_H, & \text{if } \theta_1 = \theta_H \\ 0, & \text{if } \theta_1 = \theta_L \end{cases}, \; e_2(\theta_2) = \begin{cases} e_H, & \text{if } \theta_2 = \theta_H \\ 0, & \text{if } \theta_2 = \theta_L \end{cases}. \tag{9}$$

Now let us check whether this separating strategy yields a Bayesian Nash equilibrium. Assume $e_j^*(\theta_j)$ $(j = 1, 2)$ takes this form, *i.e.*,

$$e_j^*(\theta_j) = \begin{cases} e_H, & \text{if } \theta_j = \theta_H \\ 0, & \text{if } \theta_j = \theta_L \end{cases}, \tag{10}$$

then we consider agent i's problem $(i = 1, 2, i \neq j)$. For each $\theta_i \in \{\theta_L, \theta_H\}$, agent i solves a maximization problem: $\max_{e_i} h(e_i, \theta_i)$, where by Eq (8) and Footnote 5, the object function is

$$h(e_i, \theta_i) = (w - e_i/\theta_i)P(e_i > e_j^*(\theta_j)) + (0.5w - e_i/\theta_i)P(e_i = e_j^*(\theta_j)) \tag{11}$$

We discuss this maximization problem in four different cases:
1) Suppose $\theta_i = \theta_j = \theta_L$, then $e_j^*(\theta_j) = 0$ by Eq (10).

$$\begin{aligned} h(e_i, \theta_i) &= (w - e_i/\theta_L)P(e_i > 0) + (0.5w - e_i/\theta_L)P(e_i = 0) \\ &= \begin{cases} w - e_H/\theta_L, & \text{if } e_i = e_H \\ 0.5w, & \text{if } e_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w < 2e_H/\theta_L$, then $h(e_H, \theta_L) < h(0, \theta_L)$, which means the optimal value of $e_i(\theta_L)$ is 0. In this case, $e_i^*(\theta_L) = 0$.

2) Suppose $\theta_i = \theta_L$, $\theta_j = \theta_H$, then $e_j^*(\theta_j) = e_H$ by Eq (10).

$$\begin{aligned} h(e_i, \theta_i) &= (w - e_i/\theta_L)P(e_i > e_H) + (0.5w - e_i/\theta_L)P(e_i = e_H) \\ &= \begin{cases} 0.5w - e_H/\theta_L, & \text{if } e_i = e_H \\ 0, & \text{if } e_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w < 2e_H/\theta_L$, then $h(e_H, \theta_L) < h(0, \theta_L)$, which means the optimal value of $e_i(\theta_L)$ is 0. In this case, $e_i^*(\theta_L) = 0$.

3) Suppose $\theta_i = \theta_H$, $\theta_j = \theta_L$, then $e_j^*(\theta_j) = 0$ by Eq (10).

$$\begin{aligned} h(e_i, \theta_i) &= (w - e_i/\theta_H)P(e_i > 0) + (0.5w - e_i/\theta_H)P(e_i = 0) \\ &= \begin{cases} w - e_H/\theta_H, & \text{if } e_i = e_H \\ 0.5w, & \text{if } e_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w > 2e_H/\theta_H$, then $h(e_H, \theta_H) > h(0, \theta_H)$, which means the optimal value of $e_i(\theta_H)$ is e_H. In this case, $e_i^*(\theta_H) = e_H$.

4) Suppose $\theta_i = \theta_j = \theta_H$, then $e_j^*(\theta_j) = e_H$ by Eq (10).

$$\begin{aligned} h(e_i, \theta_i) &= (w - e_i/\theta_H)P(e_i > e_H) + (0.5w - e_i/\theta_H)P(e_i = e_H) \\ &= \begin{cases} 0.5w - e_H/\theta_H, & \text{if } e_i = e_H \\ 0, & \text{if } e_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w > 2e_H/\theta_H$, then $h(e_H, \theta_H) > h(0, \theta_H)$, which means the optimal value of $e_i(\theta_H)$ is e_H. In this case, $e_i^*(\theta_H) = e_H$.

From the above four cases, it can be seen that if the wage $w \in (2e_H/\theta_H, 2e_H/\theta_L)$, then the strategy $e_i^*(\theta_i)$ of agent i

$$e_i^*(\theta_i) = \begin{cases} e_H, & \text{if } \theta_i = \theta_H \\ 0, & \text{if } \theta_i = \theta_L \end{cases} \tag{12}$$

will be the optimal response to the strategy $e_j^*(\theta_j)$ of agent j $(j \neq i)$ given in Eq (10). Therefore, the strategy profile $(e_1^*(\theta_1), e_2^*(\theta_2))$ is a Bayesian Nash equilibrium of the game induced by Γ.

Now let us investigate whether the wage $w \in (2e_H/\theta_H, 2e_H/\theta_L)$ satisfies the individual rationality (IR) constraints. Following Eq (8) and Eq (12), the (IR) constraints are changed into: $0.5w - e_H/\theta_H > 0$. Obviously, $w \in (2e_H/\theta_H, 2e_H/\theta_L)$ satisfies the (IR) constraints.

In summary, if $w \in (2e_H/\theta_H, 2e_H/\theta_L)$, then by Eq (6) and Eq (12), for any $\theta = (\theta_1, \theta_2)$, where $\theta_1, \theta_2 \in \{\theta_L, \theta_H\}$, there holds:

$$g(e_1^*(\theta_1), e_2^*(\theta_2)) = \begin{cases} (1, 0), & \text{if } \theta_1 > \theta_2 \\ (0.5, 0.5), & \text{if } \theta_1 = \theta_2 \\ (0, 1), & \text{if } \theta_1 < \theta_2 \end{cases}, \tag{13}$$

which is the social choice function $f(\theta)$ given in Eq (5). Thus, $f(\theta)$ can be implemented by the indirect mechanism Γ in Bayesian Nash equilibrium. □

5 The Bayesian implementable f is not truthfully implementable

In this section, we will show by the following proposition that a Bayesian implementable social choice function is not truthfully implementable, which means that *the revelation principle does not always hold when strategies of agents are costly.*

Proposition 4: If the misreporting cost $c_{mis} \in [0, 0.5w)$, then the social choice function $f(\theta)$ given in Eq (5) is not truthfully implementable in Bayesian Nash equilibrium.

Proof: Consider the direct revelation mechanism $\Gamma' = (\Theta_1, \Theta_2, f(\theta))$, in which $\Theta_1 = \Theta_2 = \{\theta_L, \theta_H\}$, $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$. Each agent i $(i = 1, 2)$ with private type θ_i reports a type $\hat{\theta}_i \in \Theta_i$ to the firm [6]. Then the firm performs the outcome function $f(\hat{\theta}_1, \hat{\theta}_2)$ as specified in Eq (5).

According to Proposition 1, in the direct mechanism, each agent i only reports a type and does not spend the strategic costs. The only possible cost needed to spend is the misreporting cost c_{mis} for a low-productivity agent to falsely report the high-productivity type θ_H. For agent i $(i = 1, 2)$, if his true type is $\theta_i = \theta_L$, by Eq (8) his profit function will be as follows:

$$p_i'(\hat{\theta}_i, \hat{\theta}_j; \theta_i = \theta_L) = \begin{cases} w - c_{mis}, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_L) \\ 0.5w - c_{mis}, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H) \\ 0.5w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_L) \\ 0, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_H) \end{cases}, \ i \neq j. \tag{14}$$

If agent i's true type is $\theta_i = \theta_H$, his profit function will be as follows:

$$p_i'(\hat{\theta}_i, \hat{\theta}_j; \theta_i = \theta_H) = \begin{cases} w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_L) \\ 0.5w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H), or (\theta_L, \theta_L) \\ 0, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_H) \end{cases}, \ i \neq j. \tag{15}$$

Note that the item "e_i/θ_i" occurred in Eq (8) disappears in Eq (14) and Eq (15), because each agent i does not spend strategic costs in the direct mechanism. Following Eq (14) and Eq (15), we will discuss the profit matrix of agent i and j in four cases. The first and second entry in the parenthesis denote the profit of agent i and j respectively.

Case 1: Suppose the true types of agent i and j are $\theta_i = \theta_H$, $\theta_j = \theta_H$.

$\hat{\theta}_i$ \ $\hat{\theta}_j$	θ_L	θ_H
θ_L	$(0.5w, 0.5w)$	$(0, w)$
θ_H	$(w, 0)$	$(0.5w, 0.5w)$

It can be seen that: the dominant strategy for agent i and j is to truthfully report, *i.e.*, $\hat{\theta}_i = \theta_H$, $\hat{\theta}_j = \theta_H$. Thus, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

[6] Here $\hat{\theta}_i$ may not be equal to θ_i.

Case 2: Suppose the true types of agent i and j are $\theta_i = \theta_L$, $\theta_j = \theta_H$.

$\hat{\theta}_i$ \ $\hat{\theta}_j$	θ_L	θ_H
θ_L	$(0.5w, 0.5w)$	$(0, w)$
θ_H	$(w - c_{mis}, 0)$	$(0.5w - c_{mis}, 0.5w)$

It can be seen that: the dominant strategy for agent j is still to truthfully report $\hat{\theta}_j = \theta_H$; and if the misreporting cost $0 \leq c_{mis} < 0.5w$, the dominant strategy for agent i is to falsely report $\hat{\theta}_i = \theta_H$, otherwise agent i would truthfully report. Thus, under the condition of $c_{mis} \in [0, 0.5w)$, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

Case 3: Suppose the true types of agent i and j are $\theta_i = \theta_H$, $\theta_j = \theta_L$.

$\hat{\theta}_i$ \ $\hat{\theta}_j$	θ_L	θ_H
θ_L	$(0.5w, 0.5w)$	$(0, w - c_{mis})$
θ_H	$(w, 0)$	$(0.5w, 0.5w - c_{mis})$

It can be seen that: the dominant strategy for agent i is still to truthfully report $\hat{\theta}_i = \theta_H$; and if the misreporting cost $0 \leq c_{mis} < 0.5w$, the dominant strategy for agent j is to falsely report $\hat{\theta}_j = \theta_H$, otherwise agent j would truthfully report. Thus, under the condition of $c_{mis} \in [0, 0.5w)$, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

Case 4: Suppose the true types of agent i and j are $\theta_i = \theta_L$, $\theta_j = \theta_L$.

$\hat{\theta}_i$ \ $\hat{\theta}_j$	θ_L	θ_H
θ_L	$(0.5w, 0.5w)$	$(0, w - c_{mis})$
θ_H	$(w - c_{mis}, 0)$	$(0.5w - c_{mis}, 0.5w - c_{mis})$

It can be seen that: if the misreporting cost $0 \leq c_{mis} < 0.5w$, the dominant strategy for both agent i and agent j is to falsely report, *i.e.*, $\hat{\theta}_i = \theta_H$, $\hat{\theta}_j = \theta_H$, otherwise both agents would truthfully report. Thus, under the condition of $c_{mis} \in [0, 0.5w)$, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

To sum up, under the condition of $c_{mis} \in [0, 0.5w)$, the unique equilibrium of the game induced by the direct mechanism Γ' is to fixedly report $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$, and the unique outcome of Γ' is that each agent has the same probability 0.5 to get the job offer. Consequently, the truthful report $\hat{\theta}_i^* = \theta_i$ (for all $\theta_i \in \Theta_i$, $i = 1, 2$) is not a Bayesian Nash equilibrium of the direct revelation mechanism. By Definition 23.D.3, the Bayesian implementable social choice function $f(\theta)$ given in Eq (5) is not truthfully implementable in

Bayesian Nash equilibrium under the conditions of $w \in (2e_H/\theta_H, 2e_H/\theta_L)$ and $c_{mis} \in [0, 0.5w)$, which means that *the revelation principle does not always hold when strategies of agents are costly.* □

6 Conclusions

This paper investigates whether the revelation principle holds or not when strategies of agents are costly. In Section 2, we propose the notion of profit function, and claim that *the definition of Bayesian Nash equilibrium of mechanism should be based on the profit function instead of the utility function when strategies of agents are costly.* This is the key point why the proof of revelation principle given in Proposition 23.D.1 [3] is wrong when strategies are costly.

In Section 3, we propose a simple labor model, in which agents spend strategic costs in an indirect mechanism. Section 4 and Section 5 give detailed analysis about the labor model:

1) In the indirect mechanism Γ, the profit function of each agent $i = 1, 2$ is given by Eq (8), and the separating strategy profile $(e_1^*(\theta_1), e_2^*(\theta_2))$ is the Bayesian Nash equilibrium when wage $w \in (2e_H/\theta_H, 2e_H/\theta_L)$. Thus, the social choice function f can be implemented in Bayesian Nash equilibrium.

2) In the direct mechanism, the profit function of each agent is modified from Eq (8) to Eq (14) and Eq (15). Under the condition of $c_{mis} \in [0, 0.5w)$, the unique equilibrium of the game induced by the direct mechanism is to fixedly report $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$, and the truthful report $\hat{\theta}_i^* = \theta_i$ (for all $\theta_i \in \Theta_i$, $i = 1, 2$) is not a Bayesian Nash equilibrium, which means that the revelation principle does not hold in this case.

3) Different from Kephart and Conitzer [6], the revelation principle can fail to hold even when misreporting cost $c_{mis} = 0$ (see Proposition 4).

Appendix

Proposition 23.D.1 [3]: (*The Revelation Principle for Bayesian Nash Equilibrium*) Suppose that there exists a mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ that implements the social choice function $f(\cdot)$ in Bayesian Nash equilibrium. Then $f(\cdot)$ is truthfully implementable in Bayesian Nash equilibrium.
Proof: If $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ implements $f(\cdot)$ in Bayesian Nash equilibrium, then there exists a profile of strategies $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$ such that

$g(s^*(\theta)) = f(\theta)$ for all θ, and for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i], \quad (23.D.2)$$

for all $\hat{s}_i \in S_i$. Condition (23.D.2) implies, in particular, that for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i], \quad (23.D.3)$$

for all $\hat{\theta}_i \in \Theta_i$. Since $g(s^*(\theta)) = f(\theta)$ for all θ, (23.D.3) means that, for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i], \quad (23.D.4)$$

for all $\hat{\theta}_i \in \Theta_i$. But, this is precisely condition (23.D.1) [7], the condition for $f(\cdot)$ to be truthfully implementable in Bayesian Nash equilibrium. □

Acknowledgments

The author is grateful to Fang Chen, Hanyue, Hanxing and Hanchen for their great support.

References

[1] R. Myerson, Incentive compatibility and the bargaining problem, Econometrica, vol.47, 61-73, 1979.

[2] R. Myerson, Optimal coordination mechanisms in generalized principal-agent problems, Journal of Mathematical Economics, vol.10, 67-81, 1982.

[3] A. Mas-Colell, M.D. Whinston and J.R. Green, Microeconomic Theory, Oxford University Press, 1995.

[4] Y. Narahari et al, Game Theoretic Problems in Network Economics and Mechanism Design Solutions, Springer, 2009.

[5] M. Spence, Job Market Signaling. Quarterly Journal of Economics, vol.87, 355-374, 1973.

[6] A. Kephart and V. Conitzer, The revelation principle for mechanism design with reporting cost, In Proceedings of the ACM Conference on Electronic Commerce (EC), Maastricht, The Netherlands, 2016.

[7] The condition (23.D.1) is given in Definition 23.D.3, Section 2.

[7] J. Green and J.J. Laffont, Partially verifiable information and mechanism design. Review of Economic Studies, vol.53, 447-456, 1986.

[8] L. Yu, Mechanism design with partial verification and revelation principle. Autonomous Agents and Multi-Agent Systems, vol.22, 217-223, 2011.