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

Wu, Haoyang

Wan-Dou-Miao Research Lab

11 July 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 spent in a mechanism have not been fully discussed. In this paper, we investigate the correctness of the revelation principle when strategies of agents are costly. We point out two key results: (1) The strategy of each agent in the direct mechanism is just to report a type. Each agent does not need to take any other action to prove himself that his reported type is truthful, and should not play the strategy as specified in the indirect mechanism. Hence each agent should not spend the strategic costs in the direct mechanism (see Proposition 1); (2) When strategic costs cannot be neglected in the indirect mechanism, the proof of revelation principle given in Proposition 23.D.1 of the book "A. Mas-Colell, M.D. Whinston and J.R. Green, Microeconomic Theory, Oxford University Press, 1995" 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 plays an important role in mechanism design theory [1]. According to the wide-spread textbook given by Mas-Colell, Whinston and

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

Green (Page 884, Line 24 [2]): "*The implication of the revelation principle is ... to identify the set of implementable social choice functions in 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, [3]). Related definitions about the revelation principle can be seen in Section 2, which are cited from Section 23.B and 23.D of MWG's textbook [2].

Generally speaking, agents may spend some costs in a mechanism. There are two kinds of costs possibly occurred in a mechanism: 1) *strategic costs*, which are possibly spent by agents when playing strategies [1]; and 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 [5] 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 $c(\theta, \hat{\theta}) \in \{0, \infty\}$ [6,7]. Kephart and Conitzer [5] proposed that when reporting truthfully is costless and misreporting can be 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 still holds or not when both of two kinds of costs are considered. The paper is organized as follows. In Section 2, we will analyze the strategic costs possibly occurred in an indirect mechanism, and point out two key points:
(1) The strategy of each agent in the direct mechanism is just to report a type. *Each agent does not need to take any other action to prove himself that his reported type is truthful*, and should not play the strategy as specified in the indirect mechanism. Hence each agent should not spend the strategic costs in the direct mechanism (see Proposition 1);
(2) When strategic costs cannot be neglected in the indirect mechanism, the proof of revelation principle given in Proposition 23.D.1 [2] 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 Γ, in which strategies of agents are costly. In Section 4, we prove f can be implemented by the indirect mechanism Γ in Bayesian Nash equilibrium. In Section 5, we show that the Bayesian implementable social choice function f is not truthfully implementable in Bayesian

[1] For example, agents may spend education costs in job market [4].
[2] It is usually assumed that each agent can report his true type with zero cost.

Nash equilibrium under some conditions (see Proposition 4), which contradicts the revelation principle. In the end, Section 6 draws conclusions.

2 Analysis of strategic cost

In this section, we will investigate strategic costs spent by agents in a mechanism in detail. In the beginning ,we cite some definitions from Section 23.B and Section 23.D of MWG's textbook [2]. 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 $x \in X$.

Definition 23.B.1 [2]: A *social choice function* 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 [2]: 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 $g : S_1 \times \cdots \times S_I \to X$.

Note 1: The designer has a favorite social choice function $f(\cdot)$, but he does not know the private type $\theta_1, \cdots, \theta_I$ of each agent $i = 1, \cdots, I$, and he cannot enforce agents to report their true types. What the designer can do is to construct a mechanism $\Gamma = (S_1, \cdots, S_I, g)$ to attract each agent i to voluntarily choose a strategy $s_i(\theta_i) \in S_i$ according to his own private type θ_i. After each agent i performs his strategy $s_i(\theta_i)$, the designer announces the outcome $g(s_1(\theta_1), \cdots, s_I(\theta_I)) \in X$.

Definition 23.B.5 [2]: 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 the direct mechanism, the strategy of each agent i with true type θ_i is just to choose a type $s'_i(\theta_i) \in \Theta_i$ to report, and the reported type does not need to be his true type θ_i. After the designer receives all reports, he must announce the outcome $f(s'_1(\theta_1), \cdots, s'_I(\theta_I))$.

Note 3: Strategic costs are not clearly specified in Definition 23.B.3 and Definition 23.B.5. Generally speaking, strategic costs can be financial costs or efforts spent by agents when playing strategies in a mechanism. Usually, reporting a type in the direct mechanism is considered to be costless for each agent. [3] However, strategic costs related to each agent i's strategy $s_i(\theta_i)$ in the indirect mechanism may be significant and cannot be neglected.

[3] As discussed in [5–7], misreporting in the direct mechanism may be costly.

Definition 23.D.1 [2]: 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 $\theta_i \in \Theta_i$,

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

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

Definition 23.D.2 [2]: 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 [2]: 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 $\Gamma' = (\Theta_1, \cdots, \Theta_I, f(\cdot))$. 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 $\hat{\theta}_i \in \Theta_i$.

Proposition 1: The strategy 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 the strategy $s_i^*(\theta_i)$ as specified in the indirect mechanism Γ. Hence, in the direct mechanism, each agent i should not spend the strategic costs related to strategy $s_i^*(\theta_i)$.

Proof: As pointed out in Note 2, in the direct mechanism Γ', the strategy set $S_i' = \Theta_i$, which means that the strategy s_i' of agent i with true type θ_i is just to choose a type from Θ_i to report, *i.e.*, $s_i'(\theta_i) \in \Theta_i$. Obviously in the direct mechanism, the designer still cannot enforce each agent to report truthfully, and *each agent does not need to take any action to prove himself that his reported type is truthful.* [4]
Hence, each agent i with true type θ_i will misreport another type $s_i'(\theta_i) \neq \theta_i$, $s_i'(\theta_i) \in \Theta_i$ if doing so is worthwhile. After the designer receives reports $s_1'(\theta_1), \cdots, s_I'(\theta_I)$, he has no way to verify the truthfulness of these reports.

[4] 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 is no information disadvantage from the viewpoint of the designer: the agents' types are no longer their private information, and the designer can directly specify his favorite outcome $f(\theta_1, \cdots, \theta_I)$ without any uncertainty. Note that this case contradicts the basic framework of mechanism design, therefore the assumption does not hold.

What the designer can do is just to announce the outcome $f(s_1'(\theta_1), \cdots, s_I'(\theta_I))$. Thus, *it is illegal for the designer to require each agent i play strategy $s_i^*(\theta_i)$ specified in the indirect mechanism Γ as additional actions.* As a result, in the direct mechanism Γ', each agent i with true type θ_i should not spend the strategic costs related to the strategy $s_i^*(\theta_i)$. □

Two possible questions about Proposition 1 are as follows.

Question 1: Someone may disagree with Proposition 1 and argue that Definition 23.B.5 may be modified such that each agent i also spend the strategic costs related to strategy $s_i^*(\theta_i)$. For example, the designer may define an "extended direct mechanism", in which each agent i reports a type θ_i, then the mechanism suggests each agent i to take action $s_i^*(\theta_i)$, the same as what he would take in the indirect mechanism, and the final outcome of the "extended direct mechanism" depends on both the report and the action.

Answer 1: According to the above descriptions, the input parameters of the outcome function g_e of the "extended direct mechanism" include both the reported types $\theta_1, \cdots, \theta_I$ and the actions $s_1^*(\theta_1), \cdots, s_I^*(\theta_I)$, *i.e.*,

$$g_e(\theta_1, \cdots, \theta_I, s_1^*(\theta_1), \cdots, s_I^*(\theta_I)) \in X.$$

Compared with the outcome function $g(s_1^*(\theta_1), \cdots, s_I^*(\theta_I))$ specified in the indirect mechanism Γ, the new outcome function $g_e(\theta_1, \cdots, \theta_I, s_1^*(\theta_1), \cdots, s_I^*(\theta_I))$ not only depends on each agent i's strategic action $s_i^*(\theta_i)$, but also depends on each agent i's report θ_i. Thus, the so-called "extended direct mechanism" is indeed an "extended indirect mechanism", which means that *the new mechanism not only requires agents act in the same way as what they would do in the indirect mechanism, but also requires agents announce their private types truthfully.* In comparison, in the definition of indirect mechanism, each agent i only chooses a strategic action $s_i^*(\theta_i)$, and never announces his private type. Therefore, the so-called "extended direct mechanism" contradicts the framework of mechanism design and is unreasonable. □

Question 2: Someone may object to Proposition 1 as follows: One may consider the equilibrium in the indirect mechanism, given the equilibrium, there is a mapping from vectors of agents' types into outcomes. Now suppose we take that mapping to be a revelation game, and it holds that no type of any agent can make an announcement that differs from his true type and do better.

Answer 2: To speak frankly, this argument is equivalent to the proof of revelation principle given in Proposition 23.D.1 (see Appendix). We will deeply discuss this case by the following Proposition 2.

Proposition 2: When strategic costs cannot be neglected in the indirect mechanism, the proof of the revelation principle given in Proposition 23.D.1

is wrong.

Proof: Following the description 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 $f(\theta)$ for all $\theta \in \Theta_1 \times \cdots \times \Theta_I$. For all i and all $\theta_i \in \Theta_i$,

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

for all $\hat{s}_i \in S_i$. Note that in this inequality, the utility function u_i of each agent i $(i = 1, \cdots, I)$ already reflects the fact that each agent i spends the strategic costs related to $s_i^*(\theta_i)$ specified in the indirect mechanism Γ.

Now let us take the mapping $g(s^*(\cdot))$ to be a direct revelation game, in which the strategy set of agent i is his type set, $S_i' = \Theta_i$, and the designer carries out the outcome function $g'(\cdot) = f(\cdot)$ directly. Following Proposition 1, in this direct mechanism $\Gamma' = (S_1', \cdots, S_I', g'(\cdot))$, each agent i only reports a type and does not spend the strategic costs. Thus, when strategic costs cannot be neglected in the indirect mechanism, *the utility function of each agent i in the direct mechanism Γ' should be modified from the original u_i to a new utility function u_i', which does not contain the item related to strategic costs spent by agent i in the indirect mechanism.* [5]

Consequently, in order to judge whether f is truthfully implementable or not when strategies of agents are costly, the criterion should be modified from the inequality (23.D.1) given in Definition 23.D.3 to the following new inequality (23.D.1)': f is truthfully implementable in Bayesian Nash equilibrium 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 $\hat{\theta}_i \in \Theta_i$, in which u_i' is the new utility function of agent i in the direct mechanism, and is *not equal* to u_i in the original indirect mechanism.

Therefore, the last inequality (23.D.4) in the proof of Proposition 23.D.1 is not equal to the new criterion inequality (23.D.1)'. Put differently, when strategic costs cannot be neglected in the indirect mechanism, we cannot derive that f is truthfully implementable in Bayesian Nash equilibrium by the last inequality (23.D.4). Thus, the proof of the revelation principle given in Proposition 23.D.1 is wrong. □

[5] For example, in Section 5, the utility function of agent i in the direct mechanism is modified from Eq (3) to Eq (9) and Eq (10), which does not contain the item related to the strategic costs "b_i/θ_i".

3 A labor model and a social choice function f

Here we construct a labor model which uses some ideas from the first-price sealed auction model in Example 23.B.5 [2] and the signaling model [2,4]. There are one firm and two agents. Agent 1 and Agent 2 differ in the number of units of output they produce if hired by the firm, which is denoted by productivity type. The firm wants to hire an agent with productivity as high as possible, and the two agents compete for this job offer.

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 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 an agent i $(i = 1, 2)$ of productivity type θ_i is given by a function $c(e_i, \theta_i) = e_i/\theta_i$. 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 offer with wage $w > 0$ chosen by the firm. Recall that the firm does not know the exact productivity types of two agents, but 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 $\theta = (\theta_1, \theta_2)$,

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

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

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 $b_i : \{\theta_L, \theta_H\} \to \{0, e_H\}$. The strategy set S_i is the set of agent i's all possible bids, and the outcome function g is defined as:

$$g(b_1, b_2) = (p_1, p_2) = \begin{cases} (1,0), & \text{if } b_1 = e_H, b_2 = 0 \\ (0.5, 0.5), & \text{if } b_1 = b_2 = e_H, \text{ or } b_1 = b_2 = 0 \\ (0,1), & \text{if } b_1 = 0, b_2 = e_H \end{cases}, \tag{2}$$

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

Let u_0 be the expected utility of the firm, and u_1, u_2 be the expected utilities of agent $1, 2$ in the indirect mechanism Γ respectively, then $u_0(b_1, b_2) = p_1\theta_1 + p_2\theta_2 - w$, and for $i, j = 1, 2$, $i \neq j$,

$$u_i(b_i, b_j; \theta_i) = \begin{cases} w - b_i/\theta_i, & \text{if } b_i > b_j \\ 0.5w - b_i/\theta_i, & \text{if } b_i = b_j \\ -b_i/\theta_i, & \text{if } b_i < b_j \end{cases}.$$

For the case of $b_i < b_j$, there must be $b_i = 0$. Therefore,

$$u_i(b_i, b_j; \theta_i) = \begin{cases} w - b_i/\theta_i, & \text{if } b_i > b_j \\ 0.5w - b_i/\theta_i, & \text{if } b_i = b_j \\ 0, & \text{if } b_i < b_j \end{cases}. \tag{3}$$

The item "b_i/θ_i" occurred in Eq (3) reflects the strategic costs spent by agent i of type θ_i when he performs the strategy $b_i(\theta_i)$ in the indirect mechanism. Suppose the reserved utilities of agent 1 and agent 2 are both zero, then the individual rationality (IR) constraints are: $u_i(b_i, b_j; \theta_i) \geq 0$, $i = 1, 2$.

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 (1) is Bayesian implementable, *i.e.*, it can be implemented by the indirect mechanism Γ in Bayesian Nash equilibrium.
Proof: Consider a separating strategy, *i.e.*, agents with different productivity types choose different education levels,

$$b_1(\theta_1) = \begin{cases} e_H, & \text{if } \theta_1 = \theta_H \\ 0, & \text{if } \theta_1 = \theta_L \end{cases}, \quad b_2(\theta_2) = \begin{cases} e_H, & \text{if } \theta_2 = \theta_H \\ 0, & \text{if } \theta_2 = \theta_L \end{cases}. \tag{4}$$

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

$$b_j^*(\theta_j) = \begin{cases} e_H, & \text{if } \theta_j = \theta_H \\ 0, & \text{if } \theta_j = \theta_L \end{cases}, \tag{5}$$

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: $\max_{b_i} h(b_i, \theta_i)$, where by Eq (3) the object function is

$$h(b_i, \theta_i) = (w - b_i/\theta_i)P(b_i > b_j^*(\theta_j)) + (0.5w - b_i/\theta_i)P(b_i = b_j^*(\theta_j)) \tag{6}$$

We discuss this maximization problem in four different cases:
1) Suppose $\theta_i = \theta_j = \theta_L$, then $b_j^*(\theta_j) = 0$ by Eq (5).

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

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

2) Suppose $\theta_i = \theta_L$, $\theta_j = \theta_H$, then $b_j^*(\theta_j) = e_H$ by Eq (5).

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

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

3) Suppose $\theta_i = \theta_H$, $\theta_j = \theta_L$, then $b_j^*(\theta_j) = 0$ by Eq (5).

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

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

4) Suppose $\theta_i = \theta_j = \theta_H$, then $b_j^*(\theta_j) = e_H$ by Eq (5).

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

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

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 $b_i^*(\theta_i)$ of agent i

$$b_i^*(\theta_i) = \begin{cases} e_H, & \text{if } \theta_i = \theta_H \\ 0, & \text{if } \theta_i = \theta_L \end{cases} \tag{7}$$

will be the optimal response to the strategy $b_j^*(\theta_j)$ of agent j $(j \neq i)$ given in Eq (5). Therefore, the strategy profile $(b_1^*(\theta_1), b_2^*(\theta_2))$ 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 (3) and Eq (7), the (IR) constraints are changed into: $0.5w - b_H/\theta_H > 0$. 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 (2) and Eq (7), for any $\theta = (\theta_1, \theta_2)$, where $\theta_1, \theta_2 \in \{\theta_L, \theta_H\}$, there holds:

$$g(b_1^*(\theta_1), b_2^*(\theta_2)) = \begin{cases} (1, 0), & \text{if } \theta_1 > \theta_2 \\ (0.5, 0.5), & \text{if } \theta_1 = \theta_2 \\ (0, 1), & \text{if } \theta_1 < \theta_2 \end{cases}, \tag{8}$$

which is the social choice function $f(\theta)$ given in Eq (1). 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 an example 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 $c_{mis} \in [0, 0.5w)$, the social choice function $f(\theta)$ given in Eq (1) is not truthfully implementable in Bayesian Nash equilibrium.

Proof: Consider the direct revelation mechanism $\Gamma' = (\Theta_1, \Theta_2, f(\theta))$, in which $\Theta_1 = \Theta_2 = \{\theta_L, \theta_H\}$, $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$. The timing steps of Γ' are as follows:

1) Each agent i $(i = 1, 2)$ with true type θ_i reports a type $\hat{\theta}_i \in \Theta_i$ to the firm. Here $\hat{\theta}_i$ may not be equal to θ_i.

2) The firm performs the outcome function $f(\hat{\theta}_1, \hat{\theta}_2)$ as specified in Eq (1).

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 $\theta_i = \theta_L$, his utility function will be as follows:

$$u_i'(\hat{\theta}_i, \hat{\theta}_j; \theta_i = \theta_L) = \begin{cases} w - c_{mis}, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_L) \\ 0.5w - c_{mis}, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H) \\ 0.5w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_L) \\ 0, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_H) \end{cases}, \ i \neq j. \quad (9)$$

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

$$u_i'(\hat{\theta}_i, \hat{\theta}_j; \theta_i = \theta_H) = \begin{cases} w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_L) \\ 0.5w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H), or (\theta_L, \theta_L) \\ 0, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_H) \end{cases}, \ i \neq j. \quad (10)$$

Note that the strategic cost item "b_i/θ_i" occurred in Eq (3) disappears in Eq (9) and Eq (10). Following Eq (9) and Eq (10), we will discuss the utility matrix of agent i and j in four cases. The first and second entry in the parenthesis denote the utility 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 $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

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 and Proposition 1, the Bayesian implementable social choice function $f(\theta)$ is not truthfully implementable in Bayesian Nash equilibrium under the condition of $w \in (2e_H/\theta_H, 2e_H/\theta_L)$ and $c_{mis} \in [0, 0.5w)$, which means the revelation principle does not hold. □

6 Conclusions

In this paper, we investigate whether the revelation principle holds or not when strategies of agents are costly. In Section 2, the strategic costs possibly occurred in a mechanism are analyzed in detail. The main results are Proposition 1 and Proposition 2. We point out that when strategies of agents are costly in the indirect mechanism, the utility functions of agents in the direct mechanism are different from those in the original indirect mechanism. Hence, the criterion to judge whether a social choice function is truthfully implementable should be modified as described in the proof of Proposition 2. This is the key reason why the proof of revelation principle given in Proposition 23.D.1 [2] is wrong.

In Section 3, we propose a simple labor model, in which agents spend strategic costs in an indirect mechanism. The main characteristics of the labor model are as follows:
1) Strategies of agents are costly in the indirect mechanism, *i.e.*, agent with type θ_H (or θ_L) will spend the strategic costs e_H/θ_H (or e_H/θ_L) when choosing education level e_H;
2) The productivity type of agent is his private information and not observable to the firm;
3) Misreporting a higher type is also costly, *i.e.*, a low-productivity agent can pretend to be a high-productivity agent with the misreporting cost $c_{mis} > 0$.

Section 4 and Section 5 give detailed analysis about the labor model:
1) In the indirect mechanism Γ, the utility function of each agent $i = 1, 2$ is given by Eq (3), and the separating strategy profile $(b_1^*(\theta_1), b_2^*(\theta_2))$ 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 utility function of agent is modified from Eq (3) to Eq (9) and Eq (10). 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.
Thus, the Bayesian implementable social choice function f given in Eq (1) is not truthfully implementable in Bayesian Nash equilibrium under the condition of $c_{mis} \in [0, 0.5w)$, which means that *the revelation principle does not always hold when strategies of agents are costly.*
3) Different from Kephart and Conitzer [5], this paper shows that the revelation principle can fail to hold even when misreporting cost $c_{mis} = 0$ (see Proposition 4).

Appendix

Proposition 23.D.1 [2]: (*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 equilibri-

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

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

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

[5] 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.

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

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