# Notes on Mechanism Design†

## ECON 201B - Game Theory

Guillermo Ordoñez
UCLA

February 10, 2006

## 1 Mechanism Design. Informal discussion.

Mechanisms are particular types of games of incomplete (or asymmetric) information characterized by a "principal" who would like to condition her actions on some information that is privately known by the other player (or other players), called "agent". She could simply ask the agent for his information, but he will not report it truthfully unless the principal gives him an incentive to do so, either by monetary payments or with some other instruments she controls. Since providing these incentives is costly, the principal faces a trade-off that often results in an inefficient allocation.

The principal could be the government who acts on behalf of the society, trying to achieve efficiency or a player (such as a seller) who acts on behalf of self-interest trying to maximize profits.

Examples of the first case are:

a) Regulation of a monopoly with unknown cost

b) Collection of taxes to finance public projects when the government does not know the citizens' valuations of the project

[0]† These notes were prepared as a back up material for TA session. If you have any question or comment, or notice any error or typo, please drop me a line at guilord@ucla.edu
These notes are based on Mas-Colel, Whinston and Green and Zame's lectures.

Examples of the second case are:

a) A seller (principal), not knowing the willingness to pay of the buyers (agents), need to design an auction mechanism to determine who purchase the good and the sale price

b) A second-degree price discrimination in which the monopoly (principal), who has incomplete information about the willingness to pay of the consumers (agents), designs a price schedule that determines the price to be paid by the consumer as a function of the quantity purchased (following signaling or screening processes).

c) An insurance company (principal) that designs a menu of contracts to screen customers (agents).

In an ideal world where the principal has all relevant information (including individuals' preferences), mechanism design would not be necessary. However this is not typically the case and mechanism design deals with the following question: Is it possible to achieve a particular objective by the principal in a world of selfish agents who privately enjoy some hidden information or behave in a hidden way? How?

As can be seen the origin and necessity of mechanism design relies on the existence of Bayesian games. Mechanism design is typically studied as a three step game of incomplete information, where agents' types are private information.

1) The principal designs a "mechanism" or "incentive scheme". A mechanism is a game in which the agents send costless messages, and the principal chooses an outcome or allocation based on the messages received.

2) The agents choose to accept or reject the mechanism.

3) The agents who accept the mechanism play the game specified by the mechanism.

In some cases (mainly when the principal is the government), step 2 is omitted since agents must participate in a mandatory way. Thus, participa-

tion constraints do not need to be imposed. In other cases, however, agents can freely choose whether to participate or not, for example bidders are free not to participate in an auction; buyers can refrain from buying from a firm, regulated firms can refuse to produce at all, etc.

# 2 Mechanism Design. Formal discussion.

A mechanism is composed by the following elements

- There are $I+1$ players. A "principal" $i=0$ and $I$ "agents".$i \in \{1,2,..,I\}$
- The principal does not have private information. Each agent $i$ has private information about the type $\theta_i \in \Theta_i$ that determines his preferences. The set of all possible type profile is $\Theta = \prod_{i \in \{1,2,..,I\}} \Theta_i$. Agents types are drawn from $\Theta$ according from some commonly known distribution.
- A set of outcomes $X$
- The utility function of agent $i$ with type $\theta_i$ who obtain an outcome $x \in X$ is $u_i(x, \theta_i)$
- The utility for the principal is given by $u_0(x, \theta_i)$. In the case of a social planner who tries to achieve efficiency this is directly the efficiency condition.

It is important to recall that in the case the principal knew the agents types profile $\theta = (\theta_1, \theta_2, ..., \theta_I)$, he would choose $f(\theta) \in X$. That is $f : \Theta \to X$, which is called a choice function that specifies an outcome $x \in X$ for each type profile $\theta \in \Theta$. If we're talking about a social planner, $f$ is a *social* choice function that specifies a *desirable* outcome for the whole society in the view of the social planner, $x \in X$

Naturally the problem (as in any Bayesian game) is that the principal does not know the true types of the agents, so she cannot condition her decisions on agents' types. She can only rely on messages about types sent by agents. Let $M_i$ be the set of messages agent $i$ can send such that $M = \prod_{i \in \{1,2,..,I\}} M_i$. In a mechanism these messages can be anything that you may imagine.

Hence, a mechanism is a message space $M$ and a mapping $g : M \rightarrow X$.

A **mechanism** define a Bayesian Game, and a Bayesian game is in itself a mechanism in which the messages are the actions themselves.

Let $m^* = (m_1^*, m_2^*, ..., m_I^*)$ denote an "equilibrium" of this game, being $m_i^*(\theta_i)$ the equilibrium message of agent $i$ who has a type $\theta_i$. The social choice function $f$ is said to be implementable if $g(m_1^*(\theta_1), ..., m_I^*(\theta_I)) = f(\theta)$ for all $\theta \in \Theta$

A **direct revelation mechanism** is one where each agent is asked to report his individual preferences, in which case $M = \Theta$ (and $f = g$). In an indirect mechanism agents are asked to send messages other than preferences. This difference is important because by the Revelation Principle we can focus our attention on direct revelation mechanisms.

The Revelation principle states that if a social choice function can be implemented by an indirect mechanism then it can be also implemented by a truth-telling direct revelation mechanism.

# 3 Some notions of implementability

The strongest notion of implementation is implementation in dominant strategies.

In this case $m_i^*(\theta_i)$ is the best message agent $i$ can send no matter the messages other agents sent. This is, for all $i$ and $\theta_i \in \Theta_i$

$$u_i(g(m_i^*(\theta_i), m_{-i}), \theta_i) \geq u_i(g(m_i^{'}, m_{-i}), \theta_i) \qquad \forall m_i^{'} \in M_i \text{ and } m_{-i} \in M_{-i}$$

By Revelation Principle, if $f$ can be implemented in dominant strategies, then it can be implemented by a direct revelation mechanism where truth-telling is a dominant strategy. A direct mechanism is said to be strategy-proof if revealing the true preferences is a dominant strategy for each agent and for each type $\theta_i$, i.e. each agent $i$ cannot benefit from reporting $\theta_i'$ whenever his true type is $\theta_i$. This is,

$$u_i(f(\theta_i, \theta_{-i}), \theta_i) \geq u_i(f(\theta_i', \theta_{-i}), \theta_i) \qquad \forall \theta_i' \in \Theta_i \text{ and } \theta_{-i} \in \Theta_{-i}$$

There is a weaker notion of implementability that does not require that agents reveal true types as dominant strategies.

The social choice function $f(\cdot)$ is truthfully implementable in Bayesian Nash Equilibrium if $m_i^*(\theta_i) = \theta_i$ for all $\theta_i \in \Theta_i$ and all $i$ is a BNE of the direct revelation mechanism That is, for all agents $i$ and all $\theta_i \in \Theta_i$

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)] \geq E_{\theta_{-i}}[u_i(f(\theta_i', \theta_{-i}), \theta_i)] \qquad \forall \theta_i' \in \Theta_i$$

# 4 Revelation Principle

The Revelation Principle basically tells us that anything that can be accomplished by any mechanism can actually be accomplished by a direct revelation mechanism that is individual rational and incentive compatible.

Consider any arbitrary mechanism. Each player learns his/her type and takes some action (we can think on the action itself as sending a message $m_i(\theta_i)$). Naturally, following these actions some outcome is reached.

For example, consider the case of a car that can be either good or bad, such that the seller knows the type of the car but the buyer does not. (This is the lemons example we saw in class!). A mechanism will specify that the car is sold with some probability $q$ and some amount of money $y$ is transferred between buyer and seller..

We can write $q(\sigma_B, \sigma_S|\theta)$ and $y(\sigma_B, \sigma_S|\theta)$ as the probability the car is transferred and the amount of money transferred given the buyer follows strategies $\sigma_B$ and the seller follows strategies $\sigma_S$ given the true type is $\theta$.

Suppose we have such a scheme and players follow a Bayesian Nash Equilibrium with behavioral strategies $\sigma_B^*$ and $\sigma_S^*$ respectively. We can use this equilibrium to construct a direct revelation mechanism by just setting $p(\theta) = q(\sigma_B^*, \sigma_S^*|\theta)$ and $x(\theta) = y(\sigma_B^*, \sigma_S^*|\theta)$

Basically this means the mediator learns the report of the seller and plays the game for the seller and the buyer using their equilibrium strategies and the seller's report as his type, assigning the game outcome as the outcome of the mechanism.

As should be clear by now, by construction, if the seller reports truthfully then the outcome of the mechanism will be precisely the equilibrium outcome of the original game. In particular, if in the original game both the buyer and the seller were willing to participate, the direct revelation mechanism will be individually rational. If the original game is socially efficient, so is the direct revelation mechanism

What remains to check is that, given we started our analysis considering an equilibrium for the original game, the direct revelation mechanism will be incentive compatible or not in the sense that the truth telling in the report is optimal or not (for the seller in this case).

To show this is in fact the case, fix a type $\theta$ for the seller and let $\theta'$ be the other possible type. Define a new strategy $\sigma'_S$ for the seller such that $\sigma'_S(\theta) = \sigma_S^*(\theta')$ and $\sigma'_S(\theta') = \sigma_S^*(\theta')$. In words, if the seller is of type $\theta$ he behaves as if he were type $\theta'$, otherwise he does just what he was doing before.

In the direct revelation mechanism game,

a) If the seller is of type $\theta$ and truthfully reports $\theta$, the outcome will be

$$q(\sigma_B^*, \sigma_S^*|\theta) \text{ and } y(\sigma_B^*, \sigma_S^*|\theta)$$

b) If the seller untruthfully reports $\theta'$ the outcome will be

$$q(\sigma_B^*, \sigma_S^*|\theta'), y(\sigma_B^*, \sigma_S^*|\theta') = q(\sigma_B^*, \sigma_S'|\theta), y(\sigma_B^*, \sigma_S'|\theta)$$

In other words, reporting untruthfully to the mechanism will obtain for the seller an outcome he would have obtained in the game. Because we started from an equilibrium in the original game, the seller does not wish to deviate in the game, hence he does not wish to report untruthfully to the mediator.

Hence, the mechanism is incentive compatible.

Basically the Revelation Principle tells us that, when designing a mechanism, is enough to consider direct revelation ones in which players report their types ($M_i = \Theta_i$).

If you want to get an outcome of the mechanism $f(\cdot)$ that cannot be implemented in the direct revelation mechanism, then it cannot possible be implemented in any other type of mechanism.

Given an equilibrium of any mechanism, the Revelation Principle guarantees that there is a direct mechanism for which truth telling is an equilibrium and yields exactly the same outcome.

# 5 Mediation. An example

ROW and COL are engaged in a contract dispute. They agree that COL owes money to ROW, but they disagree about how much. If they go to court, the outcome will depend on how strong each case is; the table below shows how much COL will have to pay ROW as a function of the strengths of both cases; in addition, each party will have to pay \$10 in court costs. (Numbers are expressed in thousand dollars)

| | Strong | Weak |
|---|---|---|
| Strong | \$80 | \$144 |
| Weak | \$16 | \$80 |

Each party knows whether their case is weak or strong, and believes the probability that the other's case is weak is $\frac{1}{2}$. ROW and COL ask you to mediate their dispute to avoid court. A mediation plan asks ROW and COL to report whether their cases are weak $w$ or strong $s$, and specifies, for each pair of reports $(r, c)$ a probability $Q(r, c)$ of going to court and a payment $Y(r, c)$ from COL to ROW.

a) Write down the incentive compatibility and individual rationality constraints that the mediation plan must satisfy. (Remember that going to court will result in payment from COL to ROW depending on the verdict and also court costs.)

b) Show that there is no mediation plan for which the probability of going to court is always 0.

c) If $Q(s, s) = Q(s, w) = 0$ (that is, there is no probability of going to court when ROW reports her case is strong), what is the smallest amount ROW will receive when his case is strong?

**Summarizing the game**

- COL pays to ROW as a function of the strenght of their cases.
- Cases $= \{s, w\}$
- Court cost = \$10 per each party.
- Payments of COL to ROW as in Matrix above
- $Q(r, c)$ is the probability of going to court given report $(r, c)$.
- $Y(r, c)$ is the payment from COL to ROW given report $(r, c)$.

**a) PC and IC**

**Individual Rationality Constraints (ROW player):**

These constraints must satisfy the condition that the expected utility in mediation must be higher or equal to the expected utility of going to court.

*Strong ROW*

$\text{IR}^s_{ROW}$:

$\frac{1}{2}[Y(s,s) + Q(s,s)(80-10)] + \frac{1}{2}[Y(s,w) + Q(s,w)(144-10)] \geq$

$\frac{1}{2}[80-10] + \frac{1}{2}[144-10]$

$\text{IR}^s_{ROW}$:

$\frac{1}{2}[Y(s,s) + Q(s,s)(70)] + \frac{1}{2}[Y(s,w) + Q(s,w)(134)] \geq 102$

*Weak ROW*

$\text{IR}^w_{ROW}$:

$\frac{1}{2}[Y(w,s) + Q(w,s)(16-10)] + \frac{1}{2}[Y(w,w) + Q(w,w)(80-10)] \geq$

$\frac{1}{2}[16-10] + \frac{1}{2}[80-10]$

$\text{IR}^w_{ROW}$:

$\frac{1}{2}[Y(w,s) + Q(w,s)(6)] + \frac{1}{2}[Y(w,w) + Q(w,w)(70)] \geq 38$

**Individual Rationality Constraints (COL player):**

These constraints must satisfy the condition that the expected payment in mediation must be less or equal to the expected payment of going to court.

*Strong COL*

$\text{IR}^s_{COL}$:

$\frac{1}{2}[Y(s,s) + Q(s,s)(80+10)] + \frac{1}{2}[Y(w,s) + Q(w,s)(16+10)] \leq$

$\frac{1}{2}[80+10] + \frac{1}{2}[16+10]$

$\text{IR}^s_{COL}$:

$\frac{1}{2}[Y(s,s) + Q(s,s)(90)] + \frac{1}{2}[Y(w,s) + Q(w,s)(26)] \leq 58$

*Strong COL*

$\text{IR}^w_{COL}$:

$\frac{1}{2}[Y(s,w)+Q(s,w)(144+10)]+\frac{1}{2}[Y(w,w)+Q(w,w)(80+10)] \leq$
$\frac{1}{2}[144+10]+\frac{1}{2}[80+10]$
$\text{IR}^{w}_{COL}$:
$\frac{1}{2}[Y(s,w)+Q(s,w)(154)]+\frac{1}{2}[Y(w,w)+Q(w,w)(90)] \leq 122$

**Incentive Compatibility Constraints (ROW player):**

These constraints must satisfy the condition that the expected utility of reporting the truth must be higher or equal to lie.

$$\begin{aligned}
\text{IC}^{s}_{ROW}\text{: } & \tfrac{1}{2}[Y(s,s)+Q(s,s)(70)]+\tfrac{1}{2}[Y(s,w)+Q(s,w)(134)] \geq \\
& \tfrac{1}{2}[Y(w,s)+Q(w,s)(70)]+\tfrac{1}{2}[Y(w,w)+Q(w,w)(134)] \\
\text{IC}^{w}_{ROW}\text{: } & \tfrac{1}{2}[Y(w,s)+Q(w,s)(6)]+\tfrac{1}{2}[Y(w,w)+Q(w,w)(70)] \geq \\
& \tfrac{1}{2}[Y(s,s)+Q(s,s)(6)]+\tfrac{1}{2}[Y(s,w)+Q(s,w)(70)]
\end{aligned}$$

**Incentive Compatibility Constraints (COL player):**

These constraints must satisfy the condition that the expected payment of reporting the truth must be less or equal to lie.

$$\begin{aligned}
\text{IC}^{s}_{COL}\text{: } & \tfrac{1}{2}[Y(s,s)+Q(s,s)(90)]+\tfrac{1}{2}[Y(w,s)+Q(w,s)(26)] \leq \\
& \tfrac{1}{2}[Y(s,w)+Q(s,w)(90)]+\tfrac{1}{2}[Y(w,w)+Q(w,w)(26)] \\
\text{IC}^{w}_{COL}\text{: } & \tfrac{1}{2}[Y(s,w)+Q(s,w)(154)]+\tfrac{1}{2}[Y(w,w)+Q(w,w)(90)] \leq \\
& \tfrac{1}{2}[Y(s,s)+Q(s,s)(154)]+\tfrac{1}{2}[Y(w,s)+Q(w,s)(90)]
\end{aligned}$$

**b)** $\mathbf{Q(s,s)=Q(s,w)=Q(w,s)=Q(w,w)=0}$

Imposing the restriction that all the $Q$'s (the probabilities of going to court) are equal to 0, the mechanism must satisfy the following conditions:

(1) $Y(s,s)+Y(s,w) \geq 204$
(2) $Y(w,s)+Y(w,w) \geq 76$
(3) $Y(s,s)+Y(w,s) \leq 116$
(4) $Y(s,w)+Y(w,w) \leq 244$
(5) $Y(s,s)+Y(s,w) \geq Y(w,s)+Y(w,w)$
(6) $Y(w,s)+Y(w,w) \geq Y(s,s)+Y(s,w)$

(7) $Y(s,s) + Y(w,s) \leq Y(s,w) + Y(w,w)$
(8) $Y(s,w) + Y(w,w) \leq Y(s,s) + Y(w,s)$

From (5) and (6) we have:
(9) $Y(s,s) + Y(s,w) = Y(w,s) + Y(w,w)$

And from (7) and (8) we have:
(10) $Y(s,s) + Y(w,s) = Y(s,w) + Y(w,w)$

(9) and (10) implies:
(11) $Y(s,w) = Y(w,s)$

And comparing (1) and (3) using (11), we have:
(12) $Y(s,s) + Y(s,w) \geq 204$
(13) $Y(s,s) + Y(s,w) \leq 116$

We can see from (12) and (13) that there is no mediation mechanism plan for which all the $Q$'s are equal to 0.

**c)** $\mathbf{Q(s,s) = Q(s,w) = 0}$

Imposing this restriction, the mediation mechanism must satisfy the following conditions:

(1') $Y(s,s) + Y(s,w) \geq 204$
(2') $Y(w,s) + Y(w,w) + 6Q(w,s) + 70Q(w,w) \geq 76$
(3') $Y(s,s) + Y(w,s) + 26Q(w,s) \leq 116$
(4') $Y(s,w) + Y(w,w) + 90Q(w,w) \leq 244$
(5') $Y(s,s) + Y(s,w) \geq Y(w,s) + Y(w,w) + 70Q(w,s) + 134Q(w,w)$
(6') $Y(w,s) + Y(w,w) + 6Q(w,s) + 70Q(w,w) \geq Y(s,s) + Y(s,w)$
(7') $Y(s,s) + Y(w,s) + 26Q(w,s) \leq Y(s,w) + Y(w,w) + 26Q(w,w)$
(8') $Y(s,w) + Y(w,w) + 90Q(w,w) \leq Y(s,s) + Y(w,s) + 90Q(w,s)$

From (5') and (6'), we have:

$$Y(w,s) + Y(w,w) + 6Q(w,s) + 70Q(w,w) \geq Y(s,s) + Y(s,w) \geq$$
$$Y(w,s) + Y(w,w) + 70Q(w,s) + 134Q(w,w)$$

Simplifying the expression, we have:
$0 \geq 64Q(w,s) + 64Q(w,w)$

which implies
$Q(w,s) + Q(w,w) \leq 0$

The only possible values are $Q(w,s) = Q(w,w) = 0$. We have shown in part (b) that there is no mediation mechanism in which all $Q$'s are equal to 0, as it is the case in part (c). Therefore, in the case where $Q(s,s) = Q(s,w) = 0$, in which there is no possible mediation mechanism, ROW will expect to receive the expected value of going to court ($102).