Mechanism Design

Mechanism design asks how we can provide the proper incentives to agents so that we can aggregate their preferences correctly. The mechanism design problem has been studied in economics for some time. It is also interesting to us because it maps very well to open multiagent systems with selfish agents.

Mechanism design studies how private information can be elicited from individuals. It tells us how to build the proper incentives into our protocols such that the agents will want to tell the truth about their preferences. It also tells us about the circumstances when this is impossible.

More formally, we define a *mechanism design* problem as consisting of a set of agents with the following properties:

- Each agent *i* has a *type* $\theta_i \in \Theta_i$ which is private. That is, only the agent knows its type, no one else does;
- Let $\theta = \{\theta_1, \theta_2, ..., \theta_A\}$ be the set of types;
- The *mechanism g* we are to implement will map the set of agents' actions to a particular *outcome* $o \in O$;
- Each agent *i* receives a value $v_i(o, \theta_i)$ for outcome *o*;
- The *social choice function* $f : \theta \rightarrow O$ tells us the outcome we want to achieve.

For example, the social choice function:

$$f(\theta) = \underset{o \in O}{\operatorname{argmax}} \sum_{i=1}^{n} v_i(o, \theta_i)$$

is the social welfare solution. It tries to maximize the sum of everyone's utility.

Note also that since the agent's type is usually fixed – an agent cannot change its true type, only lie about it, then v_i usually only needs to be defined for the agent's particular θ_i.

In a world with money, our mechanisms will not only choose a social alternative but will also determine monetary payments to be made by the different agents. The complete social choice is then composed of the alternative chosen as well as of the transfer of money. Nevertheless, we will refer to each of these parts separately, calling the alternative chosen the social choice, not including in this term the monetary payments.

Formally, a mechanism needs to socially choose some alternative from *A*, as well as to decide on payments. The preference of each agent *i* is modelled by a valuation function $v_i : A \to \mathbb{R}$, where $v_i \in V_i$. Throughout the rest of this section $V_i \subseteq \mathbb{R}^A$ is a commonly known set of possible valuation functions for agent *i.*

It will be convenient to use the following standard notation. Let $v = (v_1 \ldots v_n)$ be an *n*-dimensional vector. We will denote the (*n* – l)-dimensional vector in which the *i*'th coordinate is removed by $v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$. Thus we have three equivalent notations: $v = (v_1, \ldots, v_n) = (v_i, v_{-i})$. Similarly, for $V = V_1 \times \ldots \times V_n$, we will denote $V_{-i} = V_1 \times \ldots \times V_{i-1} \times V_{i+1} \times \ldots \times V_n$. Similarly we will use t_{-i}, x_{-i}, X_{-i}, etc.

A (direct revelation) *mechanism* is a social choice function $f : V_1 \times \ldots \times V_n \to A$ and a vector of payment functions $p_1 \ldots p_n$, where $p_i : V_1 \times \ldots \times V_n \to \mathbb{R}$ is the amount that agent *i* pays.

A mechanism $(f, p_1 \ldots p_n)$ is called *incentive compatible* if for every agent *i*, every $v_1 \in V_1, \ldots, v_n \in V_n$ and every $v'_i \in V_i$, if we denote $a = f(v_i, v_{-i})$ and $a' = f(v'_i, v_{-i})$, then $v_i(a) - p_i(v_i, v_{-i}) \geq v_i(a') - p_i(v'_i, v_{-i})$.

Intuitively this means that agent *i* whose valuation is v_i would prefer "telling the truth" (v_i) to the mechanism rather than any possible "lie" v'_i, since this gives it higher utility, in the weak sense.

We would like a general formula that can be used to calculate the agents' payments no matter what social choice function is given to us. Unfortunately, such a formula does not appear to exist. However, if we instead assume that the social choice function is the social welfare solution and further assume that the agents have quasilinear preferences then we can use the *Groves-Clarke mechanism* do calculate the desired payments. Agents with quasilinear preferences are those with utilities in the form $u_i(o,\theta_i) = v_i(o,\theta_i) + p_i(o)$.

Formally, the Groves-Clarke mechanism is defined as follows. If we have a social choice function:

$$f(\theta) = \underset{o \in O}{\operatorname{argmax}} \sum_{i=1}^{n} v_i(o,\theta_i)$$

then calculating the outcome using:

$$f(\tilde{\theta}) = \underset{o \in O}{\operatorname{argmax}} \sum_{i=1}^{n} v_i(o,\tilde{\theta}_i)$$

where $\tilde{\theta}$ are reported types, and giving the agents payments of:

$$p_i(\tilde{\theta}) = \sum_{j \neq i} v_j\left(f(\tilde{\theta}), \tilde{\theta}_j\right) - h_i(\tilde{\theta}_{-i})$$

where $h_i(\theta_{-i})$ is an arbitrary function, results in a strategy-proof mechanism.

Notice that the payments that *i* receives are directly proportional to the sum of everybody else's value. This is the key insight of the Groves-Clarke mechanism. In order to get the agents to tell the truth so that we may improve the social welfare we must pay the agents in proportion to this social welfare. Another way to look at it, perhaps a bit cynically, is to say that the way to get individuals to care about how everyone else is doing is to pay them in proportion to how everyone else is doing. For example, companies give shares of their company to employees in the hope that this will make them want the company as a whole to increase its profits, even if

it means they have to work longer or take a pay-cut. In effect, the Groves-Clarke mechanism places the social welfare directly into the agent's utility function.

Intuitively we would prefer that agents pay money to the mechanism, but not more than the gain that they get. Here are two conditions that seem to make sense, at least in a setting where all valuations are non-negative:

- A mechanism is (ex-post) *individually rational* if agents always get non-negative utility. Formally, if for every $v_1 \dots v_n$ we have that $v_i(f(v_1 \dots v_n)) - p_i(v_1 \dots v_n) \geq 0$;
- A mechanism has no positive transfers if no agent is ever paid money. Formally, if for every $v_1 \dots v_n$ and every *i*, $p_i(v_1 \dots v_n) \geq 0$.

The following choice of h_i's provides these two properties. The choice $h_i(v_{-i}) = \max_{b \in A} \sum_{j \neq i} v_i(b)$ is called the *Clarke pivot payment*. Under this rule the payment of agent *i* is:

$$p_i(v_1 \dots v_n) = \max_b \sum_{j \neq i} v_i(b) - \sum_{j \neq i} v_i(a)$$

where $a = f(v_1 \dots v_n)$.

Intuitively, *i* pays an amount equal to the total damage that it causes the other agents: the difference between the social welfare of the others with and without *i*'s participation. In other words, the payments make each agent internalize the externalities that it causes.

In general, Clarke-Groves mechanism charges each agent a tax, based on how much they were able to influence the decision. By informing agents of this tax mechanism in advance, they have no incentive to lie about their preferences, and therefore truth becomes the best option.

Example

The government is deciding on the number of street lights to be installed. There are 3 beneficiaries: A, B, C. There are 4 alternatives: n = 0, 1, 2, 3 where n is the number of street lights. The government's objective is to install the socially efficient number of street lights.

Total benefits and costs

Resident	No. of street lights			
	0	1	2	3
A	0	60	90	155
B	0	80	120	140
C	0	120	200	220
Cost	0	120	240	360

Net benefits with equal cost share

Resident	No. of street lights			
	0	1	2	3
A	0	20	10	35
B	0	40	40	20
C	0	80	120	100
Social net benefit	0	140	170	155

If n = 1, the total cost is 120. Hence, the cost share for each is 40. The private net benefit for A is then 60 – 40 = 20. Similarly for B and C and n = 2, 3.

Grove-Clarke taxes: person A

Resident	No. of street lights			
	0	1	2	3
A	0	20	10	35
B	0	40	40	20
C	0	80	120	100
Social net benefit	0	140	170	155

Resident	No. of street lights			
	0	1	2	3
B	0	40	40	20
C	0	80	120	100
Social net benefit	0	120	160	120

Person A is not pivotal. Without him, the net benefit is maximum at $n = 2$. With him the net benefit is maximum at $n = 2$. So his tax is 0.

Grove-Clarke taxes: person B

Resident	No. of street lights			
	0	1	2	3
A	0	20	10	35
B	0	40	40	20
C	0	80	120	100
Social net benefit	0	140	170	155

Resident	No. of street lights			
	0	1	2	3
A	0	20	10	35
C	0	80	120	100
Social net benefit	0	100	130	135

Person B however is pivotal. With him the net benefit is maximum at $n = 2$. Without him, the net benefit is maximum at $n = 3$. B's tax is therefore the difference between the sum of net benefits of others at $n = 3$ and the sum of net benefits of others at $n = 2$, i.e. $135 – 130 = 5$. B is paying the tax because his report changes the decision from $n = 3$ to $n = 2$.

Grove-Clarke taxes: person C

Resident	No. of street lights			
	0	1	2	3
A	0	20	10	35
B	0	40	40	20
C	0	80	120	100
Social net benefit	0	140	170	155

Resident	No. of street lights			
	0	1	2	3
A	0	20	10	35
B	0	40	40	20
Social net benefit	0	60	50	55

Person C is pivotal as well. With him the net benefit is maximum at $n = 2$. Without him the net benefit is maximum at $n = 1$. C's tax is therefore the sum of others' benefits at $n = 1$ and the sum of others' benefits at $n = 2$, i.e. $60 - 50 = 10$.

Net benefits with taxes

Resident	No. of street lights				Tax
	0	1	2	3	
A	0	20	10	35	0
B	0	40	40	20	5
C	0	80	120	100	10
Social net benefit	0	120	170	155	

Post tax net benefit from this scheme is 10 for *A*, 40 – 5 = 35 for B, and 120 – 10 = 110 for C.

Incentives for Truthful Revelation

Resident	No. of street lights			
	0	1	2	3
A	0	20	10	*70*
B	0	40	40	20
C	0	80	120	100
Social net benefit	0	140	170	*190*

Notice that A's net benefit is maximum at $n = 3$. Does he have an incentive to lie and change the decision to $n = 3$? Suppose A states his net benefit from $n = 3$ to be 70 instead of 35. Then, the sum of stated net benefits is maximum at $n = 3$.

Resident	No. of street lights			
	0	1	2	3
B	0	40	40	20
C	0	80	120	100
Social net benefit	0	120	160	120

But then A becomes pivotal. Without him the sum of net benefits is maximum at $n = 2$. His report changes the decision from $n = 2$ to $n = 3$. So he has to pay a tax and his tax will be equal to 160 – 120 = 40.

A's net benefit from lying will be: (net benefit from $n = 3$) – tax = 35 – 40 = –5.

A's net benefit from truthfully reporting is 10. Hence A doesn't have incentive to lie.

The same exercise can be repeated for B and C to verify that they do not have an incentive to lie either.

References

J. M. Vidal, *Fundamentals of Multiagent Systems with NetLogo examples*, http://jmvidal.cse.sc.edu/papers/mas.pdf, 2010.

N. Nisan, *Introduction to Mechanism Design (for Computer Scientists)*, in N. Nisan, T. Roughgarden, E. Tardos, V. V. Vazirani (eds.), Algorithmic Game Theory, Cambridge University Press, DOI: 10.1017/CBO9780511800481, 2007.

S. Bhattacharya, *Groves-Clarke Mechanism: An Example*, http://web.uvic.ca/~sukanta/Econ313/gcex.pdf, 2008.