Oracle-Agent Problem

10/22/2018

In this document, the goal is to write down a simple environment that yields insights into how the voting payments should be structured to ensure that the oracle reports the correct state of the world (price). In particular, it focuses on the issue of incentivizing agents to report truthfully in the face of another individual attempting to corrupt the system.

There is still some ironing out to do in the environment. As currently presented, it is not quite a fully described environment.

1 Environment

Consider a static world in which there are an exogenous number of individuals who hold margin in bi-lateral contracts. These contracts rely on a system to provide information for which each contract counter-party pays τ. There are N symmetric contracts where each counter-party holds m margin. Thus there is $2mN$ margin being held in this system.

$$
\begin{aligned}
m \equiv &\text{ margin each counterparty puts in each contract} \\
N \equiv &\text{ number of contracts in system} \\
2mN \equiv &\text{ total margin in entire system} \\
\tau \equiv &\text{ payment made by each counter-party to Oracle} \\
T \equiv &\text{ total payments made to Oracle from contract market } = 2\tau N
\end{aligned}
$$

There is a random state of the world $S \in \{0, 1\}$. There is an agent, "the Oracle", who is in charge of ensuring that the system provides correct information about S. There is also a malevolent agent who would like the system to provide false information about S. The malevolent agent is a counter-party in each of the contracts, and, if the wrong state gets reported, can collect the other party's margin. This means the profit from corrupting the system is $\text{PFC} = mN$. Note: This is a worst case analysis. This is the largest incentive that one could have for corrupting this system.

$$
\begin{aligned}
S \equiv & \text{ state of the world} \\
\text{PFC} \equiv & \text{ Profit from corruption } = mN \\
\text{CoC} \equiv & \text{ Cost of corruption}
\end{aligned}
$$

The Oracle sells the right to vote on what S is to a measure one of agents who can purchase this right to vote at price p. Both the Oracle and malevolent agent can provide conditional payments to the voters in order to entice them to tell the truth (or lie). The Oracle, can only condition the payments it makes on an individual's action and the actions made by all other individuals, i.e. $\varepsilon(x_i, x^{-i})$. However, the malevolent agent can condition on an individual's action, the actions made by other individuals, and S, i.e. $\hat{\varepsilon}(x_i, x^{-i}, S)$. Note, the oracle will pay individuals in "rights to vote for tomorrow". These rights to vote for tomorrow have value $p' = p$ if the Oracle announces the correct state and $p' = 0$ if the Oracle is corrupted.

$$
\begin{aligned}
p \equiv & \text{ Cost of purchasing a right to vote} \\
p' \equiv & \text{ Value of the right to vote for tomorrow} \\
\varepsilon(x, x^{-i}) \equiv & \text{ The payment the Oracle makes to voter} \\
\hat{\varepsilon}(x, x^{-i}, S) \equiv & \text{ The payment the malevolent agent makes to voter} \\
\text{PFC} \equiv & \text{ Profit from corruption } = mN \\
\text{CoC} \equiv & \text{ Cost of corruption } = \int_i \hat{\varepsilon}(x, x^{-i}, S)
\end{aligned}
$$

Each of the voting individual knows the state S and maximizes their utility given by:

$$
V = \max_{\text{No Vote,Vote}} \{0, \max_{x \in \{0,1\}} E\left[p' \varepsilon(x, x^{-i}) + \hat{\varepsilon}(x, x^{-i}, S)\right] - p\}
$$

The Oracle chooses $\varepsilon(x, x^{-i})$ to solve

$$
\begin{aligned}
V^O = \min_{\varepsilon(x, x^{-i})} & \int_i \varepsilon(x_i, x^{-i}) di \\
& \text{subject to} \\
& p' \int_i \varepsilon(x_i, x^{-i}) di \leq T + \int_{i \in \text{Voters}} p di \quad \text{(Budget Constraint)} \\
& V(S) \geq V(1 - S) \quad \text{(Incentive Compatible)}
\end{aligned}
$$

The malevolent agent chooses $\hat{\varepsilon}(x, x^{-i}, S)$ in (an attempt?) to solve

$$V^M = \min_{\hat{\varepsilon}(x,x^{-i},S)} \int_i \hat{\varepsilon}(x_i, x^{-i}, S)di$$

subject to

$$\int_i \hat{\varepsilon}(x_i, x^{-i}, S)di \leq mN \quad \text{(Budget Constraint)}$$

$$V(1-S) > V(S) \quad \text{(Lie Compatiblity)}$$

Obviously both of these problems can't be solved. They need to be connected such that the Oracle is minimizing the cost of maintaining the Oracle truthful subject to the fact that the malevolent agent is attempting to corrupt the system... If the malevolent agent weren't there then the Oracle would simply set $\varepsilon(x, x^{-i}) = 1$ and people would report the truth because they would be indifferent. What makes the problem interesting is that the malevolent agent is "tempting" the agents to report falsely.

1.1 Solving the Model

We want to solve two problems:

1. Find optimal payment scheme $\varepsilon(x, x^{-i})$... As written this isn't well specified... Need to sort out what exactly this looks like.
2. Next step would be to take a parameterized approximation to $\varepsilon(x, x^{-i})$ and solve for the Ramsey solution in a dynamic versoin of this world

1.2 Next Steps

Based on the learnings from solving the static world exercise, we (hopefully) can use the same functional form for the ideal (Mirrlees) payout function to extend this model to the dynamic world and solve for the simpler Ramsey problem of the optimal flat tax rate for the dynamic system.