Bribery Attack Notes

UMA Project

January 20, 2020

1 Overview

2 Basic bribery attack setting

One of the pieces of feedback that we have received is that the UMA Protocol is exposed to the simplest form of bribery. For the sake of clarity, we have tried to transform the arguments made by others into a formal game theory model.

2.1 Formal setting

We consider the setting for the data verification machine (DVM) under which the system issues π tokens to any individual who votes with the majority. The objective of the DVM is to provide accurate information by incentivizing people to vote "correctly." We normalize the tokens and the rewards such that they are expressed in dollar terms. Note we assume that if the system becomes corrupted then the tokens are worth 0 in dollar terms.

There is a continuum (of measure one) of agents who can either vote to corrupt or to not corrupt. There exists a 3rd party individual who would like to corrupt the system, they offer a contract which rewards any agent who votes to corrupt $\tilde{x}$ (in dollars).

We can then determine the payoffs for the action taken by each individual:

- If the system is corrupted and the individual voted to corrupt, they receive $\tilde{x}$
- If the system is corrupted and the individual voted to not corrupt, they receive 0
- If the system is not corrupted and the individual voted to not corrupt, they receive $1 + \tilde{x}$
- If the system is not corrupted and the individual voted to not corrupt, they receive $1 + \pi$

2.2 Equilibrium description

We consider the space of all mixed strategy Nash equilibrium (MSNE). That is, given a strategy p, which denotes the probability with which each individual will vote to not corrupt, there must not be any profitable deviation on the individual level.

An individual's payoff is given by

$$V = p(\mathbb{I}_{\text{corrupt}} \cdot 0 + \mathbb{I}_{\text{not corrupt}} \cdot (1 + \pi)) + (1 - p)(\mathbb{I}_{\text{corrupt}} \cdot \tilde{x} + \mathbb{I}_{\text{not corrupt}} \cdot (1 + \tilde{x}))$$

Suppose $\exists\ p > 0.5$ such that p is a MSNE then it must be that

$$\begin{aligned} p(1+\pi) + (1-p)(1+\tilde{x}) &> (1+\tilde{x}) \\ p((1+\pi) - (1+\tilde{x})) &> 0 \\ (1+\pi) &> (1+\tilde{x}) \\ \pi &> \tilde{x} \end{aligned}$$

However, if $\pi > \tilde{x}$ then the only p that is a MSNE is $p = 1$ which implies all individuals vote to not corrupt.

Now suppose $\exists\ p < 0.5$ such that p is a MSNE then it must be that

$$\begin{aligned} p(0) + (1-p)\tilde{x} &> 0 \\ \tilde{x} &> 0 \end{aligned}$$

However, if $p < 0.5$ and $\tilde{x} > 0$ then the only MSNE is $p = 0$ which implies that all individuals vote to corrupt.

This demonstrates how we might be exposed to certain forms of bribery — In the case in which $\tilde{x} > \pi$ the only MSNE is $p = 0$ which implies certain corruption of the DVM.

3 Proportional reward bribery attack setting

While the bribery model in the previous section is useful in highlighting where the system might be vulnerable, it is missing some important features of the actual system that help secure the system against such attacks. Namely, rewards are proportional to the number of voters in the majority.

3.1 Formal setting

We consider a similar setting to Section 2.1. There is a DVM that would like to reveal the truth by incentivizing people to vote "correctly." It does this by offering a total amount of rewards π split among all individuals who vote with the majority. We continue to normalize the token value and rewards to dollar amounts and work with the assumption that corruption results in the vote token being valued at 0.

There is a continuum (of measure one) of agents who can either vote to corrupt or to not corrupt. There exists a third party individual who would like to corrupt the system, they offer a contract which rewards any agent who votes to corrupt with $\tilde{x}$ dollars.

We can determine the payoffs for the action taken by each individual as a function of the fraction of individuals who vote to not corrupt p.

- If the system is corrupted and the individual voted to corrupt, they receive $\tilde{x}$
- If the system is corrupted and the individual voted to not corrupt, they receive 0
- If the system is not corrupted and the individual voted to not corrupt, they receive $1 + \tilde{x}$
- If the system is not corrupted and the individual voted to not corrupt, they receive $1 + \frac{1}{p}\pi$

3.2 Equilibrium description

We continue to use MSNE as our equilibrium concept. Rewards are now written as

$$V = p(\mathbb{I}_{\text{corrupt}} \cdot 0 + \mathbb{I}_{\text{not corrupt}} \cdot (1 + \frac{1}{p}\pi)) + (1-p)(\mathbb{I}_{\text{corrupt}} \cdot \tilde{x} + \mathbb{I}_{\text{not corrupt}} \cdot (1 + \tilde{x}))$$

Suppose $\exists\ p > 0.5$ such that p is a MSNE then it must be that

$$\begin{aligned} p(1 + \frac{1}{p}\pi) + (1-p)(1+\tilde{x}) &> 1 + \tilde{x} \\ p((1 + \frac{1}{p}\pi) - (1+\tilde{x})) &> 0 \\ \pi &> p\tilde{x} \end{aligned}$$

This is quite similar to the outcome of last time. However, it's important to note that in order to successfully convince the "$0.50 + \varepsilon$" person, it will require that $\tilde{x}$ is twice as high as in the previous case. This results in a multiplicity of equilibria in which $\tilde{x} > \pi$, but the system does not become corrupted.

In the case of $p < 0.5$ we see much of the same math because people have already assumed that the system will be corrupted — If people already know the system will be corrupted then it is a self-fulfilling prophecy.

If we instead consider a trembling hand equilibrium, we introduce additional motive for individuals to That being said, there are additional equilibria with p close to 0.5 and they each require that $\tilde{x}$ be higher it becomes even more difficult to corrupt the DVM