Mechanism Design

(Based on Shoham and Leyton-Brown (2008). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, Cambridge.)

Leen-Kiat Soh

Agents are **self-interested**. Each is motivated to maximize its utility

If voting non-truthfully allows a *rational* agent to benefit *more* than voting truthfully, it will vote *falsely*

A MAS designer ***cannot*** tell whether an agent is voting truthfully or not

However, is it possible to design the system such that the agents are *motivated* to tell the truth because it is the more beneficial thing (rationally) for them to do?

Introduction

- ***Mechanism design*** is a **strategic** version of social choice theory, which adds the assumption that **agents will behave so as to maximize their individual payoffs**
  - Social choice theory is *nonstrategic*; it takes the preferences of the agents as given, and investigates ways in which they can be aggregated
- a.k.a. **implementation theory** *or* **"inverse game theory"**

Example | Strategic Voting

- **Babysitting Example**. Four kids: Will, Liam, Vic, and Ray. Three choices: (a) going to the video arcade, (b) playing basketball, and (c) going for a leisurely car ride (c). The activity with the highest number of votes, with ties broken alphabetically, will be selected.
- The true preferences of the kids are:

$$Will\colon\ b \succ a \succ c$$
$$Liam\colon\ b \succ a \succ c$$
$$Vic\colon\ a \succ c \succ b$$
$$Ray\colon\ c \succ a \succ b$$

How can we design the MAS such that Ray will be motivated to choose to vote truthfully?

- Consider that Will, Liam, and Vic vote according to their true preferences
- **If Ray knows the other three's true preferences prior to the vote, how should he vote?**
  - Telling the truth: If he votes for *c*, the winner is *b*.
  - Not telling the truth: If he votes for *a*, the winner is *a* (tie-breaking alphabetically).
  - Which one is better to Ray, *a* or *b*?
  - Strategically, Ray should vote for *a*: **Ray has motivation to "lie"**

Game Theory: Given an interaction among a set of agents, how do we **predict or prescribe the course of action** of the various agents participating in the interaction?

Mechanism Design: Given certain **desired behaviors on the part of agents** and ask **what strategic interaction** among these agents might **give rise** to these behaviors

Purpose

- We will assume **unknown individual preferences**, and ask whether we can design a game such that, **no matter what the secret preferences** of the agents actually are, **the equilibrium of the game is guaranteed to have a certain desired property or set of properties**
  - Engineering emergent behavior, or
  - Incentive engineering
- **The most famous application of mechanism design is *auction theory* (Chapter 11)**

Connection to MAS?

Mechanism design is perhaps the most "computer scientific" part of game theory, since it concerns itself with **designing effective protocols for distributed systems.**
The key difference from the traditional work in distributed systems is that **in the current setting the distributed elements are *not* necessarily cooperative**, and **must be motivated to play their part**.

Silly Question: **An ice cream store wanted to find out which flavors of ice cream that its customers love the most, in order to come up with a Top-3 list. What if they used this strategy to motivate customers to vote: "Voting is open for 100 days. After that, we will sell the top flavor ice cream at 25% of its current price, the second-place ice cream at 50% of its current price, and the third-place ice cream at 75% of its current price." Would you vote truthfully? If not, how should the ice cream store revise its "incentive"?**