Robustness Evaluation of Incentive Mechanisms (Extended Abstract)

Yuan Liu
School of Computer Engineering
Nanyang Technological University, Singapore
yliu3@ntu.edu.sg

Jie Zhang
School of Computer Engineering
Nanyang Technological University, Singapore
zhangj@ntu.edu.sg

ABSTRACT

A general assumption for incentive mechanisms is that all agents are rational and seek to maximize their utility. When some agents are irrational and launch various attacks, these mechanisms may fail to work. To address the issue of evaluating the robustness of incentive mechanisms, we propose a robustness metric in this paper. It is inspired by the studies of the evolutionary game theory and defined as the maximum percentage of irrational agents existing in the system while it is still better off for rational agents to perform desired strategies. Then a simulation framework is designed to measure the robustness of incentive mechanisms, and is verified to be able to produce the same results as those by theoretical analysis. Finally, we demonstrate the usage of our simulation framework in evaluating and comparing the robustness of two incentive mechanisms where irrational agents adopt different attacking strategics.

Categories and Subject Descriptors

I.2.11 [**ARTIFICIAL INTELLIGENCE**]: Distributed Artificial Intelligence ÍC Intelligent agents, Multiagent systems

General Terms

Experimentation, Performance

Keywords

Mechanism Design, Irrational Agents, Robustness

1. ROBUSTNESS METRIC

Incentive mechanisms have been proposed to address the free-riding problem and incentivize agents to perform strategies desired by mechanism designers. For example, in reputation systems, information shared by agents is aggregated to model targets' reputation. Thus, truthful information is desired, but agents may keep silent or provide misleading information [3]. Side-payment mechanisms [1] aim to promote truthful information where honestly reporting is a Nash equilibrium strategy. A general assumption for incentive mechanisms is that every agent is rational and has the belief that others are also rational. However, this assumption may not always be true in real life as irrational agents often exist and launch various attacks to the system, causing rational agents performing undesired strategies [2]. Therefore, a critical issue is how to evaluate the robustness of incentive mechanisms against irrational agents, which has not been addressed in the literature.

We start by looking into the evolutionary game theory [2] where the assumption that agents are rational is relaxed. An important concept in the theory is the evolutionary stable strategy where there exists a small θ such that agents still prefer to adopt the strategy given $x < \theta$ deviating agents (invaders) in the population. What we are concerned with is the maximum value of θ, referred to as the robustness of incentive mechanisms. Formally, it can be defined as follows:

Definition (**Robustness**) The robustness of an incentive mechanism is the maximum proportion of irrational agents which mutate their strategy in a certain way (referred to as a type of attacks) such that all rational agents still sustain the strategy desired by the incentive mechanism.

2. SIMULATION FRAMEWORK

Two challenges are imposed on evaluating the robustness of an incentive mechanism through theoretical analysis: 1) attacks launched by irrational agents may be too complex to be theoretically modeled. For example, some attacks are combinations of different types of attacks; 2) the settings of incentive mechanisms may be too complex to be theoretically abstracted. Therefore, we propose a simulation framework to measure the robustness of incentive mechanisms. We verify that our framework can produce the same results as those by theoretical analysis on simple games and attacks.

Our simulation framework is based on the evolutionary process [2] to study the strategy dynamics of a specific population and effectively model the evolution of strategy propagation regardless of the implementation of incentive mechanisms and attacks. It is outlined in Procedure 1.

In the framework, we gradually involve more irrational agents until all rational agents abandon the desired strategy of the incentive mechanism. The fitness of a strategy (Line 9) is reflected by the expected payoff agents can obtain by performing the strategy. The mapping between payoff and fitness is captured by the intensity of selection. Given the fitness of strategies, the probability of a strategy being selected for reproduction (Line 10) is proportional to the fitness of the strategy and the number of agents performing the strategy. The mutation rate of an agent from one strategy to another is determined by the Fermi function [2] (Line 12).

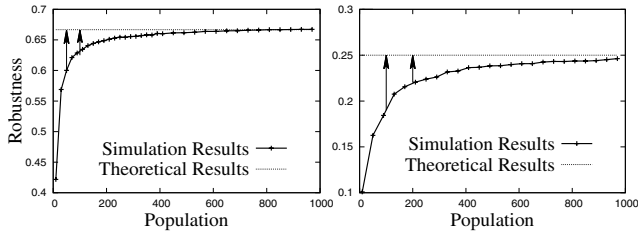
Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.*

As the evolutionary process (Lines 8-12) involves much randomness, we introduce a small probability $\epsilon = 0.05$ to indicate the extent to which rational agents sustain or abandon the desired strategy (Lines 13-15).

```
1  Implement the incentive mechanism;
2  Set initial number of irrational agents N = 1;
3  Set maximum generations G to a large number;
4  while true do
5      Set M rational agents, perform desired strategy;
6      Involve N irrational agents with a certain attack;
7      for i = 1 → 10000 do
8          for g = 1 → G do
9              Calculate the fitness for each strategy;
10             Reproduce an agent with chosen strategy;
11             Randomly choose an agent to die;
12             Agents mutate their strategies;
13     Calculate probability P that rational agents
       sustain the desired strategy;
14     if P < 1 − ε then
15         Break;
16     else
17         N++;
18 Output the robustness R = (N−1)/(M+N−1);
```

Procedure 1: The Simulation Framework

We verify the simulation framework on a symmetric coordinate game where each agent chooses its action between Left and Right. For any two agents in the game, each agent can only gain some payoffs if both of them choose the same action, for example, x payoffs if both choose Left or y payoffs if both choose Right. Thus, two Nash equilibriums exist in the game. Assume that {Left, Left} is the desired equilibrium, then rational agents would always choose Left if other agents are also rational. Irrational agents, instead, would choose any other strategies but not Left. If irrational agents always take Right, then the robustness can be calculated as the proportion of the irrational agents such that the expected payoff of Left equals to Right, which is $\frac{x}{x+y}$. We simulate the game in our framework by setting different values for x and y. The results in Figure 1 show that the robustness measured by our simulation framework can always converge to the theoretical results, which validates the effectiveness of the simulation framework.

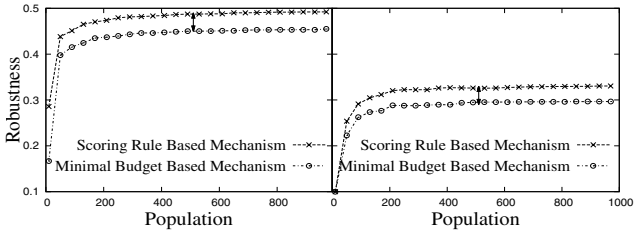


(a) x=1, y=0.5 (b) x=0.33, y=1

Figure 1: Robustness of the Simulated Games

3. EVALUATE INCENTIVE MECHANISMS

We also demonstrate the usage of our simulation framework in evaluating and comparing the robustness of two side-payment incentive mechanisms (a scoring rule based and a minimal budget based) [1] against irrational agents performing two kinds of common attacks (constant and collusive). In e-marketplaces employing side-payment incentive mechanisms, binary ratings (1 or 0) are reported by buyers about sellers. A rating r_i from a buyer b_i about a seller will be compared with a rating r_j issued by another buyer b_j (called reference reporter) about the same seller. Then, buyer b_i can gain the payoff $\pi(r_i|r_j)$. Specifically, in the scoring rule based mechanism, $\pi(1|1) = 1$, $\pi(1|0) = \pi(0|1) = 0$, and $\pi(0|0) = 1$. In the minimal budget based mechanism, $\pi(1|1) = 0.086$, $\pi(1|0) = \pi(0|1) = 0$, and $\pi(0|0) = 0.1$. Irrational buyers performing constant attacks will always provide untruthful ratings to all sellers. Irrational buyers performing collusive attacks collude to provide untruthful ratings towards a group of (50%) sellers. In the simulations, sellers are implemented as rational and seek to maximize their profit. Rational buyers choose the sellers with the highest reputation (represented by the average of received ratings) to conduct transactions with.



(a) Constant Attack (b) Collusive Attack

Figure 2: Robustness of Incentive Mechanisms

The results in Figure 2 show that the scoring rule based incentive mechanism has higher robustness than the minimal budget based incentive mechanism. The reason is that the latter sacrifices the robustness property in order to achieve the minimal side-payment imposed on the e-marketplace owner. In addition, the side-payment mechanisms bear lower robustness against collusive attacks than constant attack, indicating that they are less robust when irrational agents launch more complex attacks.

4. CONTRIBUTIONS AND FUTURE WORK

The contributions of our current work include: 1) a robustness metric for evaluating incentive mechanisms with the existence of irrational agents; 2) a simulation framework based on the evolutionary process to measure and compare the robustness of incentive mechanisms. For future work, we will demonstrate the usage of our framework in evaluating other incentive mechanisms against various sophisticated attacks launched by irrational agents.

5. ACKNOWLEDGEMENTS

This work is supported by the MOE AcRF Tier 1 Grant (M4010265 RG15/10) awarded to Dr. Jie Zhang.

6. REFERENCES

[1] R. Jurca and B. Faltings. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34(1):209–254, 2009.

[2] S. Saha and S. Sen. Predicting agent strategy mix of evolving populations. In *Proceedings of AAMAS*, 2005.

[3] J. Zhang. *Promoting Honesty in Electronic Marketplaces: Combining Trust Modeling and Incentive Mechanism Design*. PhD thesis, U of Waterloo, 2009.