1 Reputation Systems

1.1 Review

1.1.1 Key Points

1. **Reputation systems** are (online) mechanisms that aggregate feedback from users’ past experiences, to enable more informed decisions of other users in the future.

2. **Moral hazard** can arise in any bilateral transaction, where an agent has the opportunity to deviate from the promised course of action, but where the action he takes is hidden such that he doesn’t bear the full (negative) consequences/risk of the action that he takes. If the moral hazard problem is not addressed and the percentage of cheaters becomes too large, a market might break down completely, because the risk for the buyers might become too large.

3. **Adverse selection** arises in situations with asymmetric information between the agents involved in a transaction. If the information asymmetries are large enough that adverse selection becomes a big problem, legitimate sellers would be deterred from entering the market, leaving only sellers with inferior products.

4. **Modeling reputation systems through game theory**: a group of \( n + 1 \) agents with \( n \) odd, where in each round agents are paired up randomly to play the Prisoners’ Dilemma game. How do the strategies for two-player PD translate to this setting? Now assume there is a public counter documenting how many times a player has defected. How can this reputation system affect the equilibria?

5. **Whitewashing attacks** consist of an agent exiting the system when she has a bad reputation and creates a new identity to start anew. This can be combated by raising the bar to entry, for instance requiring new users to verify accounts with a photo ID.

6. **eBay’s old reputation system** had a flaw that made submitted feedback viewable immediately, even if the other party had yet to leave their own feedback. In this system, sellers could manipulate buyers to leave positive feedback, making the reputation feedback uninformative. Potential fixes investigated by eBay included simultaneous revealing of feedback and detailed seller ratings, the latter of which eBay decided to implement.
7. **eBay’s new reputation system** Included one-directional feedback: the buyer can provide feedback on several attributes of the interaction with the seller, and this information is delayed and pooled so that the seller doesn’t know which buyer provided the specific feedback. In a later change, sellers are prevented from giving negative feedback (they can only give positive feedback or no feedback).

### 1.1.2 Things you should be able to do

1. Understand moral hazard and adverse selection problems in different settings.

2. Reason about the strategic aspects of the reputation game.

3. Discuss eBay’s old and new reputation system and explain the pros and cons of different reputation mechanisms.

4. Discuss the different design decisions (directionality, type of feedback requested, etc.) of various reputations systems such as those of Amazon, Airbnb, and Yelp

### 1.2 Exercises

1. Whitewashing

   Consider the repeated Prisoner’s Dilemma game where agents have a published reputation of how many times they have played D. Once an agent plays D twice, everyone else begins defecting against him. Assume that agents are allowed to whitewash their identity and re-appear as a new agent. Let $f$ denote the initiation fee paid in round 0 and every subsequent re-entry into the market in the repeated Prisoner’s Dilemma game.

   What is the sequence of payoffs you’d get with

   (a) defecting once and then cooperating forever?

   (b) defecting forever?

   (c) defecting twice, whitewashing, and then repeating?

2. Reputation

   (a) Comment on eBay’s old reputation system. What were some pros and cons of this system?

   (b) Discuss the pros and cons of a blind feedback system (i.e. the one eBay considered implementing).

   (c) Discuss the pros and cons of the detailed seller rating system (i.e. the one eBay implemented).

   (d) Amazon has one mechanism for rating sellers and another for rating items whether it be a book or a new gardening tool. What are some pros and cons of the item rating systems, where agents assign a score (1 to 5 stars) and may leave a comment about items that they have ordered?
(e) Yelp has a similar star rating system for restaurants, where an agent does not have to prove he actually attended the restaurant. What are some pros and cons of this system that differentiate it from a rating system where only users who’ve ordered the product can rate it?

2 Information Elicitation

2.1 Review

1. Scoring rules attempt to incentivize agents to report their beliefs about some event/outcome that will be realized in the future.

2. An agent’s payment (which can be negative) is a function of his report and the realized outcome.

3. A scoring rule is strictly proper if it incentivizes an agent to report his true belief.

4. The quadratic and logarithmic scoring rules are examples of strictly proper scoring rules.

5. The logarithmic scoring rule is
\[ t_{\log}(q, o_k) = \ln(q_k). \]

6. The quadratic scoring rule is
\[ t_{\text{quad}}(q, o_k) = 2q_k - \sum_{k'=0}^{m-1} q_{k'}. \]

7. If \( t(q, o_k) \) is a strictly proper scoring rule on \( m \) outcomes, then any rule \( t'(q, o_k) = \alpha_k + \beta \cdot t(q, o_k) \) obtained via a positive affine transform with \( \alpha \in \mathbb{R}^m \) and \( \beta \in \mathbb{R}_{>0} \) is strictly proper.

8. Sometimes agents must exert effort to figure out their beliefs. By increasing the value of \( \beta \) in our scoring rule, we can make it worthwhile for agents to exert effort.

9. In the peer prediction method, agents are asked to report their signal and payments are made based on joint reports of the signals. The agents report signals without knowing what the other agents’ reports are.

10. Two simple mechanisms are output agreement and 1/Prior. There are also mechanisms based on scoring rules.

   a. Output Agreement
   \[ t_i(r_1, r_2) = 1(r_1 = r_2) \text{ for } i \in \{1, 2\} \]
   for reports \( r_1, r_2 \) from agents 1 and 2, respectively

   b. 1/Prior
   \[ t_i(r_1, r_2) = \frac{1}{P(r_1)} 1(r_1 = r_2) \text{ for } i \in \{1, 2\} \]
   for reports \( r_1, r_2 \) from agents 1 and 2, respectively

   c. Multi-task Bonus-Penalty Mechanism
   \[ t_i(r_1, r_2) = 1(r_1^{(0)} = r_2^{(0)}) 1(r_1^{(1)} = r_2^{(2)}) \text{ for } i \in \{1, 2\} \]
   for reports \( r_1, r_2 \) on a set of three tasks from agents 1 and 2, respectively
11. One issue with running peer prediction mechanisms in practice is that there can be additional, uninformative equilibria. Participants in a peer-prediction mechanism may coordinate and receive high payments without revealing any useful information. A good solution is to use the multi-task bonus/penalty mechanism (and its generalizations).

2.2 Exercises

2.2.1 Biased Coin

Suppose we have a biased coin that comes up heads with an unknown probability $p$. Based on our observations of past coin flips, we form the belief that $p$ is \( \frac{\text{number of } H}{\text{number of flips}} \).

We have flipped the coin 9 times and have received 3 heads.

1. Under a logarithmic scoring rule, what should I report as my belief about the probability that the next flip is an H to maximize expected score?
2. What is my expected payoff from reporting truthfully?
3. What is my expected payoff from reporting 0.25?
4. What is my expected payoff from reporting 0.5?

2.2.2 Peer Prediction

Suppose we have true, hidden states of the world $H = \{0, 1\}$ and possible signals $S_i = \{0, 1\}, \forall i \in \{1, 2\}$, and that $P(H = 0) = 0.4$. Further suppose that $P(X_i = 0|H = 0) = 0.6$ and $P(X_i = 1|H = 1) = 0.8$.

1. Calculate the joint distribution on signals.
2. Calculate the payments in truthful equilibrium under Output Agreement. Is OA strictly proper here? Justify your answer.
3. Are there additional, uninformative equilibria under Output Agreement? If so, what are they?
4. Now think about the scoring-rule based peer-prediction mechanism. Calculate payments under the logarithmic scoring rule for each pair of reports. Then, give transformed payments such that strict properness still holds but payments are between 0 and 1.
5. Calculate payments under the $1/Prior$ peer prediction mechanism. Is this mechanism strictly proper? Justify your answer.