Manipulation-Resistant Reputation Systems
Eric Friedman, Paul Resnick and Rahul Sami

Rachael Colley

University of Amsterdam, ILLC
rachael@colley.co.uk

October 22, 2018
Overview

- The effects of reputations via the Prisoners Dilemma
- Whitewashing
- The Peer Prediction Model and strategies.
- Transitive Trust
- Incentives for Honest Reporting
- Sybil Attacks
- Extensions of what we’ve seen
Why are Reputation systems important?

The internet

- Example: An online shopper might get a better price online from a small unknown seller which could lead to being defrauded or low-quality items.
- We want to know how reliable an agent is.
- A history may reveal information about the entity’s ability.
- A history may create an incentive to perform reliably.
### Example (The Prisoners Dilemma)

<table>
<thead>
<tr>
<th></th>
<th>Player B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Player A</td>
<td>(1, 1)</td>
</tr>
<tr>
<td></td>
<td>(2, −1)</td>
</tr>
</tbody>
</table>

- If A and B were to play this once then \((D, D)\) is a dominant strategy.
- If it were played infinitely many times then players may want to take a lower payoff in one round to ensure a higher payoff in a later round.
The discounted payoff to player $i$ in stage $t$ of the game is:

- The discounted payoff to player $i$ in stage $t$ of the game is $\pi^t_i(\delta)^t$, where $\pi^t_i$ is player $i$’s payoff in round $t$ and the discount factor $0 \leq \delta < 1$.
- The discount factor $\delta$ reflects the preference of getting a payoff in the current round rather than in later rounds.

Discounted average payoff of a strategy (played infinitely):

$$\overline{\pi_i} = (1 - \delta) \sum_{t=0}^{\infty} (\delta)^t \pi^t_i$$
The Prisoner’s Dilemma- repeated play

Consider the following strategy

*Grim*: Play $C$ unless any player has played $D$ in the previous round. If the players strategy was (Grim, Grim), then a player $i$’s discounted average payoff would be:

$$\bar{\pi}_i = (1 - \delta)(1 + \delta \cdot 1 + \delta^2 \cdot 1 + ...) = 1$$

(Given $0 \leq \delta < 1$)
The Effect of Reputations

Theorem

The strategy \((Grim, Grim)\) is a Subgame Perfect Nash Equilibria (SPNE) when \(\delta \geq \frac{1}{2}\).

Proof.

We only need to consider ‘single deviations’ (deviating from \(Grim\) once and then returns to play it again).

If \(A\) deviated in the first round and played \(D\), then all future rounds both players always play \(D\). \(A\)’s discounted average payoff will be:

\[
\pi_i' = (1 - \delta)(2 + \delta \cdot 0 + \delta^2 \cdot 0 + \ldots) = 2(1 - \delta)
\]

As \(\pi_i' = 2(1 - \delta)\) and \(\pi_i = 1\), it is only not worth deviating if \(1 \geq 2(1 - \delta)\). Hence, \(\delta \geq \frac{1}{2}\).
The Effect of Reputations

Theorem

*The strategy (Grim, Grim) is a Subgame Perfect Nash Equilibria (SPNE) when* \( \delta \geq \frac{1}{2} \).

Remarks

We can see here that if \( \delta \) is small, then the promise of future payoffs is not sufficient to constrain players behaviour.

Other strategies

Lets look at some ‘less grim’ strategies. Instead of punishing from the point of defection, only punishing for a set number of time steps. This could lead to a higher cooperative equilibrium for higher values of \( \delta \).
Now we can extend this game to many players, consider $N + 1$ players, where $N$ is odd. At each round each player is paired up at random to play the prisoners dilemma.

**Strategy**

- **Reputational-grim**: each agent begins with a ‘good’ reputation and keeps it if she plays $C$ with those who have a good reputation and $D$ with those who do not.
- If all players play the strategy reputational-grim, this is also a SPNE.
- This is because the punishment is the same as in the full Grim strategy.
Whitewashing

Definition (Whitewashing)
When an agent has a bad reputation, we say that they are ‘whitewashing’ when they re-register under a new username.

In the previous example if an agent played $D$ and got a higher return, then they could re-register to get a clean slate and would not have deal with the repercussions.
Whitewashing

How to prevent whitewashing

- **Initiation fee**: Paying an upfront cost of $f$ to register, this will prevent whitewashing as long as $f$ is sufficiently high. This is not always possible or feasible, for example online services such as ebay.

- **Verify the identity**: The organisation could force members to reveal their true names, e.g. airbnb requiring proof of identity.

- **A new strategy**: ‘Pay your dues’ (PYD) Agents play $C$ against any veteran who has never deviated from PYD, otherwise play $D$ against the veteran. Play $D$ against a newcomer, unless you are a newcomer too, then play $C$. This leads to the most socially efficient SPNE.
Problems

Two problems with the previous examples is that:

- Agents need form and report opinions which takes time and effort, even though it benefits others.
- The second challenge is honesty. How do you get agents to report without fear of retaliation.

Solution

Creating an explicit reward system for honest rating. There isn’t always objective information, so we might look to compare agents reports and reward agreement.

We are going to look at a model which aims to create a peer prediction method.
The Model

- The set of raters $I$, such that $|I| \geq 3$ and can be (countably) infinite.
- The product has type $t$, this is its quality. $t = 1, \ldots, T$ ($T$ finite and the same for all raters)
- The probability of getting a certain type is common knowledge. Such that $P(t) > 0$ for all $t$ and $\sum_{t=1}^{T} P(t) = 1$
- The raters can have a perception of the product, we will call the signal. The possible signals are $S = \{s_1, \ldots, s_M\}$. We denote the the signal received by rater $i$ as $S^i$.
- The conditional probability of receiving a signal given a type is common knowledge and identical for each rater. $f(s_m|t) = P(S^i = s_m|t)$ and $f(s_m|t) > 0$ for all $m$ and $t$. 
The Model

Strategies

- This is a simultaneousness reporting game, hence there needs to be an announcement from the centre about the players.
- \( x = (x_1, \ldots, x^I) \) is the vector of the announcements from all of the raters in a particular instance.
- Let \( \bar{x}^i = (x^i_1, \ldots, x^i_M) \) be the set of announcements rater \( i \) will give any signal she receives. Where \( x^i_m \in S \).
- \( \tau_i(x) \) is the transfer paid to the rater \( i \) when \( x \) has been announced by the raters. The vector representing the total transfers paid is \( \tau(x) = (\tau_1(x), \ldots, \tau_I(x)) \).
An announcement strategy $\overline{x}^i$ is a best response to $\overline{x}^{-i}$ for a player $i$, if for each $m$:

$$\forall \hat{x}^i \in S \quad \mathbb{E}_{S^{-i}}[\tau_i(\hat{x}^i_m)|S^i = s_m] \geq \mathbb{E}_{S^{-i}}[\tau_i(x^i_m)|S^i = s_m]$$

$\overline{x}$ is a Nash Equilibrium if $\overline{x}^i$ is a best response for all raters $i$.

This is a Nash Equilibrium when for all raters $i$ and for all observations $m$ we have that $x^i_m = s_m$. 

\textbf{Definition (Truth Revelation)}

Rachael Colley (ILLC)
The Model

Example

- The product can be of type $H$ or of type $L$.
- The raters can perceive the product as $h$ or $l$.
- We assume that $f(h|H) = 0.85$, $f(h|L) = 0.45$, $f(l|H) = 0.15$ and $f(l|L) = 0.55$.
- Consider the transfer function which gives a positive point to those who are in the majority and no point to those in the minority.
- Consider the strategies $((h, l), (h, l))$ this leads to $P(agree) = 0.625$ and $P(disagree) = 0.375$.
- Consider the strategy $((h, h), (h, l))$, this leads to $P(agree) = 0.65$ and $P(disagree) = 0.35$.
- Not better off by playing truthfully.
Transitive trust

- We don’t have objective feedback
- Now with transitive trust the credibility of an agents is determined by the credibility of the raters of that agent.

**Definition (Credible agent)**

If an agent $i$ has positive feedback from other currently credible agents, then $i$ should be included in the set of credible agents.

This is a recursive construction for which we need to define a way to compute the credible calculations. We can now represent this in graphical form.
At a given point of time we can have a trust graph $G = (V, E, t)$

- $V$ is the set of agents.
- $E$ is the set of directed edges between agents.
- $t : E \rightarrow \mathbb{R}^+ \setminus \{0\}$ are weights on the edges.
- $t(i, j)$ denotes the summary feedback or trust that $i$ reports about $j$ based on their interactions. $t : E \rightarrow \mathbb{R}^+ \setminus \{0\}$.

**Definition (Reputation function of $G$)**

The function $F$ defined as such $F : G \rightarrow \mathbb{R}^{|V|}$, provides a rank of the agents.

We let $F_v(G)$ denote the reputation value of a vertex $v$. 
Transitive Trust: Example

PageRank

- $v \in V$ is a web page.
- $(v, w) \in E$ is a directed graph showing that $v$ has a hyperlink to page $w$.
- $t(v, w) = \frac{1}{Out(v)}$, where $Out(v)$ is the outdegree of $v$.
- The rank function of this can be given by:

$$F_v(G) = \epsilon + (1 - \epsilon) \sum_{v' :(v', v) \in E} F_{v'}(G) t(v', v)$$

- See example.
Incentives for honest reporting

In these models there might even be an incentive to report dishonestly to influence what others think about her. Hence, we want a reputation function where an agent can’t improve her own standing by strategically giving feedback.

**Definition (Rank-strategyproof)**

If for every graph $G$ and every agent $v \in V$, an agent $v$ cannot boost her rank ordering by strategic choice of how she rates other agents.

**Remark**

PageRank is not rank-strategyproof.
Definition (A Sybil Attack)

A single agent creates many fake online identities to boost its primary online identity. Formally:

- A node can create any number of ‘Sybil nodes’ with any set of trust values between them.
- The node can divide the incoming trust edges amongst the Sybil nodes in any way to preserve the total trust \( \sum_{v' \mid (v', v) \in V} t(v', v) \).
Definition (Sybil Strategy)

Given a graph $G = (V, E, t)$ and a user $v \in V$, we say that a graph $G' = (V', E', t')$ along with a subset $U' \subseteq V'$ is a sybil strategy for user $v$ in the network $G = (V, E, t)$ if $v \in U'$ and collapsing $U'$ into a single node with the label $v$ in $G'$ yields $G$. We can refer to $U'$ as the sybils of $v$, and denote the sybil strategy by $(G', U')$.

Example (PageRank)
Now we want to define two notions of ‘sybilproofness’

**Definition (Value-sybilproof)**

A reputation function $F$ is value-sybilproof if for all graphs $G = (V, E)$ and all users $v \in V$, there is no sybil strategy for $v$, $(G', U')$ with $G = (V', E')$ such that for some $u \in U'$, $F_u(G') > F_v(G')$.

**Definition (Rank-sybilproof)**

A reputation function $F$ is rank-sybilproof if for all graphs $G = (V, E)$ and all users $v \in V$, there is no sybil strategy $(G', U')$ for $v$ with $G = (V', E')$ such that for some $u \in U'$ and $w \in V \setminus \{v\}$ that $F_u(G') \geq F_w(G')$ whilst $F_v(G) < F_w(G)$.
**Definition (Symmetric)**
The reputation of the nodes depends only on the structure of the graph.

**Theorem**
There is no (non-trivial) symmetric rank-sybilproof reputation function.

**Proof.**
- Given a graph $G = (V, E)$ and a reputation function $F$, let $v, w \in V$ with $F_w(G) > F_v(G)$.
- Consider $G'$ which is simply two disjoint copies of $G$, where $U$ is the second copy of $G$ with $v$.
- By symmetry, there is a node $u \in U$ such that $F_w(G') = F_u(G')$.
- Therefore, $F$ is not rank-sybilproof.
Sybil Attacks

In the previous proof we allowed for a node to create a new copy of the graph as sybil nodes. We can restrict this.

**Definition (K-rank-sybilproof)**

A reputation function is K-rank-sybilproof if it is rank-sybilproof for all possible sybil strategies \((G', U')\), with \(|U'| \leq K + 1\)

**Theorem**

*There is no symmetric K-rank-sybilproof non-trivial reputation function for \(K > 0\).*
### Max-flow reputation function

\[ F_v(G) \] is the maximum flow from some start node \( v_0 \) to \( v \). 
\( v_0 \) has a privileged position, usually given the a ‘trusted’ node.

### Theorem

*The max-flow based ranking mechanism is value-sybilproof*

### Proof.

- This proof follows from the max-flow min-cut theorem.
- All of the sybil’s of \( v \) and \( v \) must be on the same side of the cut.
- Therefore, they are on the other side of the cut to the source \( v_0 \).
- Thus, no sybil can have a value higher than the min-cut, which is equal to \( F_v(G) \).
- See previous example.
Extensions of what we’ve seen

- There has been a lot of study into *distributed reputation systems* where the reputations are computed by the users themselves.
- We want robustness against attack but we also want a functioning reputation system. In a system where all agents had zero reputation would be perfectly secure but completely useless. Hence, there has been a lot of study into finding a balance between these factors.
- One of the main challenges of reputation systems is trying to put all aspects we looked at coherently together.
Conclusion

We saw:

- multiple rounds of the prisoner’s dilemma and the strategy *Grim*. When the discount factor $\delta \geq \frac{1}{2}$ this lead to a SPNE.
- an extension of this linking our game to the reputation of agents.
- the attempt at whitewashing by agents to improve their reputations and possible ways to get around this.
- the need to move to a peer-prediction method to promote honesty.
- the model of peer-prediction and the example with the products $H$ and $L$.
- the graphical presentation of transitive trust and the application to PageRank.
- Sybil attacks and how this can lead to the notion of strategies being ‘sybilproof’.