Reputation Systems

Seminar about algorithms 2013/2014

Adam Furmańczuk

adam.furmanczuk@fu-berlin.de

Prof. Dr. Wolfgang Mulzer
Email: mulzer@mi.fu-berlin.de
Yannik Stein
Email: yannik.stein@fu-berlin.de
Theoretical Computer Science
Computer Science Institute
Department of Mathematics and Computer Science
Freie Universität Berlin
http://www.inf.fu-berlin.de/inst/ag-ti/

Table of contents

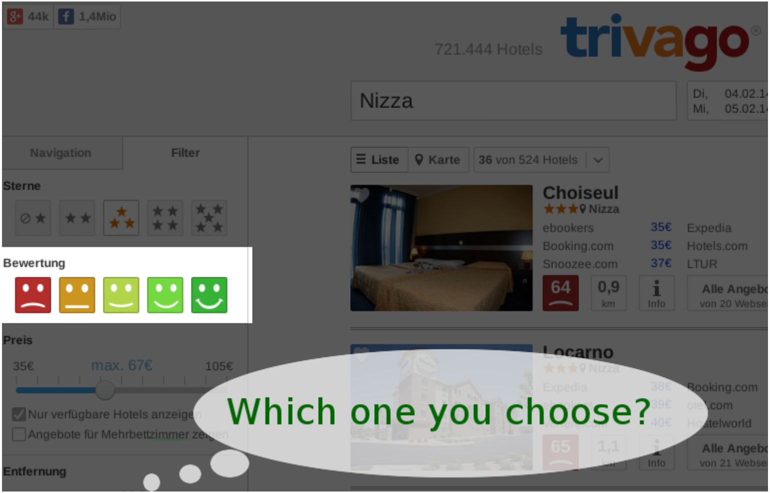
Freie Universität Berlin

Motivation

- Why reputation matters?
  - good rep $\rightarrow$ trust in ability and reliability in:
    - business (e-commerce, internet auctions)
    - education (researchers involve you more often in publications)
    - online communities (stackoverflow, specialized message boards..)
  - for technical matters (e.g. : rank peers in p2p systems)
  - act strategically thinking about history we generate
- active research topics at our university

Motivation

- It affects your holidays:



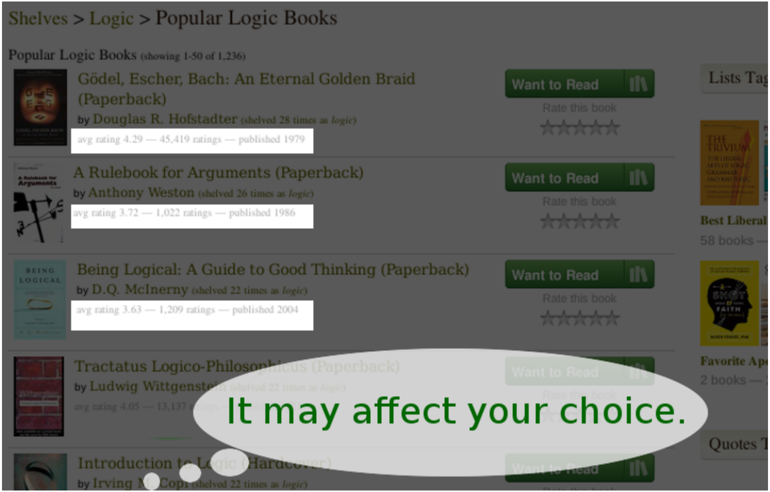
Motivation

- It affects where you buy:

Motivation

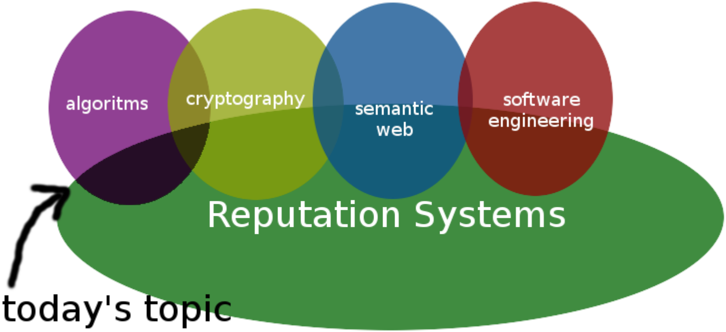
Freie Universität Berlin

- It influences your thinking:



Reputation Systems

- Where do we research reputation systems?



- recall Nadja's talk on game strategy:

Tab. 2.1: Das Gefangenendilemma

		Wesson	
		Geständnis	Schweigen
Smith	Geständnis	$(-4,-4)$	$(0,-10)$
	Schweigen	$(-10,0)$	$(-2,-2)$

- let's modify, so crime pays:
  - both cooperate: each gain 1
  - both cheat: each gain 0
  - one cooperates: cheater gain 2, other lose 1

Grim

- Let's play the **grim strategy** in an infinite model:
  - coopeate unless any player has cheated in previous round
  - both play grim $\rightarrow$ **Subgame Perfect Nash Equilibrium** (SPNE)
  - payoff at stage i: π_i^t
  - discount factor δ : $0 \leq \delta \leq 1$
  - discounted avg payoff: $\bar{\pi}_i = (1 - \delta) \sum_{t=1}^{\infty} \delta^t \pi_i^t$
- **proof**:
  - consider single cheating at t=0:
  - $\rightarrow$ (cheat, cheat) for remaining roounds
  - avg payoff: $(1 - \delta)(2 + \delta * 0 + \delta^2 * 0 + ...) = 2(1 - \delta)$
  - when cooperating: $(1 - \delta)(1 + \delta * + \delta^2 + ...) = 1$
  - cheating not advantegous when $\delta \geq \frac{1}{2}$
  - same argument for any period $t > 0$

Reputational Grim

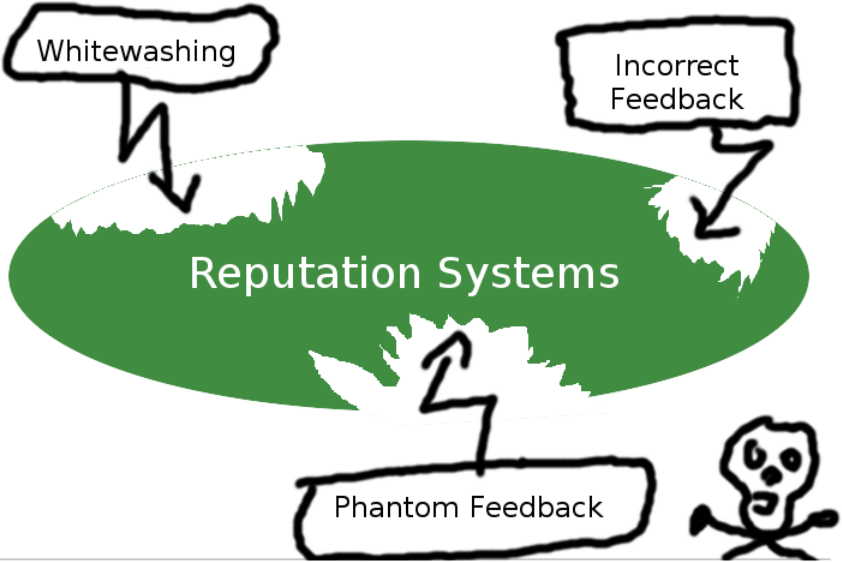
- now N players (N - is even) get paired up at random
- each player starts with good rep
- may keep good rep when:
  - cooperate with good rep platers
  - cheat on players with bad rep
- when played by all players $\rightarrow$ SPNE for $\delta \geq \frac{1}{2}$
- **proof**:
  - for cheater punishment is same as in full grim strategy

Personalized Grim

Freie Universität Berlin

- history remembered but not shared
- player view game as seperated into N unrelated games
- SPNE for $\delta \geq 1 - \frac{1}{2N}$

Threats to reputation systems:



Whitewashing:

Freie Universität Berlin

- starting afresh with new pseudonym
- threat to previous games: cheat until bad rep and start afresh
- not desired

Whitewashing prevention:

- simple mean: use initiation fee *f* (upfront cost)
  - When is cheating to expansive?
  - cheating payoff:
    $\bar{\pi}' = (1-\delta)(2 - f + \delta * (1-f) + \delta^2 + \delta^3 + ...) = 2(1-\delta)$
  - rep grim payoff: $\bar{\pi} = (1-\delta)(1 - f + \delta + \delta^2 + ...)$
  - for SPNE one needs $\bar{\pi} \geq \bar{\pi}' \rightarrow f \geq \frac{1}{\delta}$

Whitewashing prevention: Pay your dues (PYD)

- allow *veterans* to cheat against *newcommers*
- PYD is most efficient equilibrium
- whitewashers are treated same as newcommers
  - assume αN real newcommers arrive every period
  - when there are more αN newcomers, whitewasher present $\rightarrow$ extremely noise influenced (fragile)
  - add "noise"to the model allowing *veteran* accidentally ($\epsilon > 0$) play D and return as whitewasher

Whitewashing prevention: other

- other strategies to prevent whitewashing include:
  - newcomers play only against newcomers
  - reveal true identities (Post-ident, etc)

Incorrect Feedback

- underprovision (not enough feedback available)
- dishonest or distorted feedback

Prevent Incorrect Feedback

- create incentive by rewarding feedback
- naive solution: compare reports to peers and reward agreement
  - problem: herding and information cascade
- **peer prediction method**

Modelling peer prediction

Freie Universität Berlin

- Assumptions in model *truthful revelation*:
  - product quality constant (observed with error)
  - *rater* sends evaluation of product to *center*
  - *center* awards and punishes *raters* based on their msgs
  - no independent information available
  - *raters* risk neutral, try to maximize wealth
- each *rater* has perception of product (signal)
- **no information on other *raters* signal**

Simultaneous reporting game

- *center* asks *rater* to announce signal
- after receiving all signals information is shared
- *center* computes *transfers*

Simultaneous reporting: modelling details

Freie Universität Berlin

- $t = 1, ..T$ indexed types to be rated
- set of *raters* I; $|I| \geq 3$
- signals:
  - $S = \{s_1, ..., s_M\}$ possible signals
  - S^i random signal received by *rater i*
  - signals distributed by:
    $f(s_m \mid t) = Pr(S^i = s_m \mid t)$;
    $\forall s_m, t : f(s_m|t) > 0 \wedge \sum_{m=1}^{M} f(s_m|t) = 1$
  - conditional distribution different for each type, e.g. :

	h	l
H	$f(h \mid H) = 0.85$	$f(l \mid H) = 0.15$
L	$f(h \mid H) = 0.55$	$f(h \mid H) = 0.45$

$\rightarrow Pr(h) = 0.5 * 0.85 + 0.5 * 0.45 = 0.65$

Simultaneous reporting: more details

- announcements by *center*:
  - single announcment: $x^i \in S$
  - announcements for all *raters*: $x = \{x^1, ..., x^I\}$
  - *rater i's* announcment fo signal s_m: $x^i_m \in S$
  - *rater i's* announcment strategy: $\bar{x}^i = (\bar{x}^i_1..\bar{x}^1_M)$
  - announcment strategy vector: $\bar{x} = (\bar{x}^1, .., \bar{x}^I)$
  - strategy vector without *rater i*: $\bar{x}^{-i}$

- transfer:
  - transfer paid to *rater i* when announcing x: $\tau_i(x)$
  - transfer paid to all *raters*: $\tau(x) = (\tau_1(x), ..., \tau_I(x))$

Simultaneous reporting: observations

- $\bar{x}^1$ **best response to** $\bar{x}^{-1}$**, if for each m**:
  $\forall \hat{x}^i \in S : E_{S^{-1}}[\tau_i(\bar{x}^i_m, \bar{x}^{-1}) \mid S^i = s_m] \geq E_{S^{-1}}[\tau_i(\hat{x}^i, \bar{x}^{-1}) \mid S^i = s_m]$

- $\bar{x}$ is Nash Equilibrium (NE) for *simultaneous reporting* if formula holds for $i = 1..I$ and strict NE if inequality is strict.
- $\rightarrow$ *truthful revelation* is NE of *simultaneous reporting* , if formula holds for all i when $x^i_m = s_m$ for all i and m

Scoring in simultaneous reporting:

Freie Universität Berlin

- score rule T *strictly proper*, if rater maximizes expected score by announcing true belief
  - e.g.: logarithmic scoring rule:
    - penalize player the log of probability s/he assigned to occurred event
  - assign reference rater $r(i) \neq i$, strict NE when:
    - assuming r(i) reports honestly $x^{r(i)}(s_m) = s_m$
    - S^i stochastically informative for $S^{r(i)}$
    - since r(i) honest $\rightarrow$ stochastically informative on r(i) as well
    - for any $S^i = s^*$, rater chose $x^i \in S$ to maximize: $\sum_{m=1}^{M} T(s^{r(i)} \mid x^i) Pr(S_{r(i)} = s_n \mid S_i = s^*)$
    - T strictly proper $\rightarrow$ formula maximized for $x^i = s^*$
    - $\rightarrow$ **given r(i) truthful, i's best option truthful as well**

- **however: *truthful reporting* not unique equilibrium**, e.g. :
  - report h all the time
  - report l all the time

Transitive trust

- problem: lack of objective feedback $\rightarrow$ agents rep is tied to non feedback action $\rightarrow$ calculated from other agents rep
- initial trusted set (agents trusting each other) $\rightarrow$ recursive calc
- problem when to propagate trust and when to stop calculation

Transitive trust - modelling

- view rep systems as trust graphs $G = (V, E, t)$:
  - V set of agents
  - E set of directed edges
  - $t : E \to \mathbb{R}^+$
  - $F : G \to \mathbb{R}^{|V|}$
  - $F_v(G)$ reputation value of $v \in V$

- simple PageRank: $F_v(G) = e + (1 - e) \sum_{v' | (v', v) \in E} F_v(G) t(v', v)$
- max flow: $F_v(G)$ maximum flow from some start node $v_0 \in V$ to v
  - select one or several trusted start nodes
- Pathrank: $F_v(G)$ shortest path to some $v_0 \in V$, distance is inverse of trust value
- **Problem: choose rep function so v cannot boost rep with strategic feedback**

Phantom attack

- also called sybil attack, sock puppet:
  - create many fake ids to boost rep of primary id
- sybil strategy:
  - given $G = (V, E, t)$ create $G' = (V', E', t')$ with $U' \subseteq V'$ if $v \in U'$ and collapsing U' to v in G' results in G

Sybil-proof

- value sybil proof function F:
  - $\forall G \wedge v \in V$: if no sybil strategy exists for any $u \in U'$ that would fullfil $F_u(G') > F_v(G)$.
- rank-sybil proof function F:
  - $\forall G \wedge v \in V$: if no sybil strategy exists: $u \in U' \wedge w \in V \backslash \{v\} \ : \ F_u(G') \geq F_w(G') \wedge F_u(G) < F_w(G)$
  - $\rightarrow$ **no symmetric sybil attack proof function F can exist**
- max flow is value sybil-proof:
  - max flow equals min-cut $\rightarrow$ all sybils of v on same side of cut and other side than $\rightarrow$ no sybil can have higher value than v
- **max flow is not rank sybil proof:**
  - reduce values for nodes where it is on max-flow path:

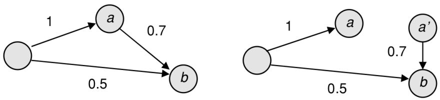


Figure 27.2. Node (a) improves its ranking by adding a sybil (a') under max-flow.

Sybil-proof continued

- **pathrank is value and rank sybil proof**
  - v cannot increase its shortest path $\rightarrow$ value proof
  - v could affect w if v on shortest path, but $F_v(G) > F_w(G)$ $\rightarrow$ rank-proof

Summary

Freie Universität Berlin

- reputation systems are in everyday use
- because of popularity under constant attack
- three weaknesses of reputation systems:
  - whitewashing
  - incorrect feedback
  - phantom feedback
- can be solved with algorithms

Perspective

Freie Universität Berlin

- many open problems remain:
  - settings with lack of honest third party (compare *center* from simultaneous reporting game)
  - users able to intercept each other msgs
  - $\rightarrow$ distributed reputation systems
- analyse google‘s or yahoo‘s Trust Rank algorithm
- design (and implement) system robust to described attacks

Sources

- Nisan, Roughgarden, Tardos, Vazirani. Algorithmic Game Theory, chapter 27