A Fresh Look at Trust and Reputation Systems

Christopher J. Hazard

cjhazard@ncsu.edu

North Carolina State University

February 13, 2009

Trust & Reputation

- What is Trust?
  - presumption of fulfilled action
  - assured reliance of character, ability, strength, or truth (Merriam-Webster)
- What is Reputation?
  - Belief that something is a certain way

On Computational Trust...

- "Never trust a computer you can't throw out a window." - Steve Wozniak

Trust Within Autonomous Agents

- Many applications
  - automated procurement, web services, recommender systems, personal assistant agents
- Trust research spans disciplines
  - Will you buy food from company X?
  - Are you telling the truth?
- Even within Computer Science
  - No common definition
  - No common metrics to compare one system to another
  - No common criteria or desiderata

Contribution:

- A set of common dimensions to categorize trust systems
- A set of common desiderata for building trust systems
- A set of common metrics to compare trust systems
- Results comparing 5 widely cited models, and one new model...

Outline

Trust Meta-Survey

- Ramchurn Huynh Jennings '04 (RHJ)
- Artz Gil '07 (AG)
- Sabater Sierra '05 (SS)
- Jøsang Ismail Boyd '07 (JIB)
- Dellarocas '06 (D)
- Mui Halberstadt Mohtashemi '02 (MHM)
- Commonalities between surveys

Common Dimensions Overview

- Incentive Compatibility (RHJ, D)
- Access v Action (RHJ, AG, JIB)
- Focus on Adverse Selection (SS, JIB, D, RHJ)
- Focus on Moral Hazard (SS, JIB, D, RHJ)
- Context Dependency (SS, JIB, MHM AG)
- Aggregation Breadth (RHJ, JIB, MHM, AG, D)

Dimension: Incentive Compatibility

- Incentive compatibility: honesty is rational
- If reputation is primary mechanism, then usually no.
  - e.g. eBay
- If incentive compatible mechanism, then yes.
  - e.g. Fly on a commercial arline - buy ticket first

Dimension: Access v Action

- Access Trust
  - Identity & Permissions
  - Security & encryption domain
  - Enables action trust
  - e.g. Account for online banking, Kerberos
- Action Trust
  - Provision, delegation, reciprocation, good-faith, etc.
  - e.g. eBay, Epinions
  - Focus of remainder of classification

Dimension: Focus on Adverse Selection

- Intrinsic quality: fixed ability/attribute
- Reliability, collaborative filtering
- Cause: information asymmetry, cure: signalling
- Often with infrequent interaction
- Can measure with statistics, but caveats
- e.g. Epinions, Jøsang '98

Dimension: Focus on Moral Hazard

- Moral Hazard: whether to uphold standards or promises
- Cause: rationalism, cure: sanctioning
- Often with frequent interaction
- Cannot measure by standardly applying statistics
- e.g. Contrite tit-for-tat (Sudgen '86, Boyd '89)
- Few systems focus only on moral hazard

Notes on Adverse Selection and Moral Hazard

- Completely independent dimensions
- Found together in most real-world environments
- Dual meanings of subjective
  - Qualified, affective
  - Relative to self (moral hazard)
- Objective is either
  - Mesurable
  - Global metric (adverse selection)

Dimension: Context Dependency

- Number of different dimensions of reliability measures used
- Examples:
  - Subjective (affective): 0
  - Probability of positive interaction (Jøsang '98): 1
  - Discount factor & reliability (Smith & desJardins '09): 2
  - Video game review (graphics, sound, gameplay, etc.): 4
  - Review of a manufacturer's product lineup: N

Dimension: Aggregation Breadth

- Individual accumulation (decentralized) v global reputation (centralized)
- Prejudice, priors, & credentials
- e.g. eBay v Netflix v Lone observations (Sen '02)

Aggregation Mechanism

- Closely coupled with Aggregation Breadth
- Supported by JIB
- Popular methods
  - Summation (eBay)
  - Bayesian (Jøsang '99, Hazard '08)
  - Discrete values (Cognitive approaches)
  - Belief models (Yu & Singh '02)
  - Fuzzy models (Sabater & Sierra '01)
  - Flow models (Pagerank, Eigentrust)

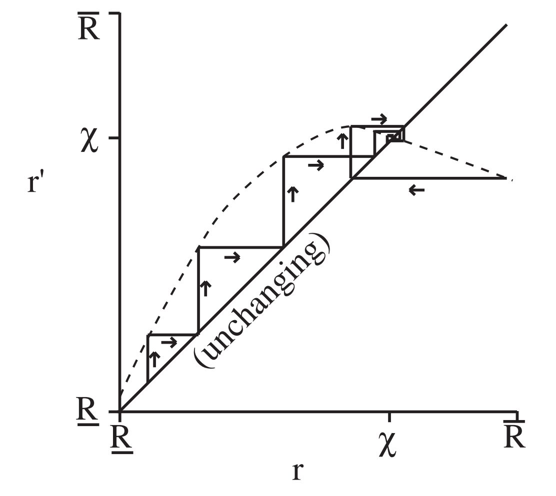
Trust System Desiderata Overview

- Evidential (adverse selection, moral hazard)
- Aggregable (adverse selection, aggregation breadth)
- Viable/tractable
- Robust (moral hazard)
- Flexible (combine info from contexts)
- Privacy enhancing (collection minimization)

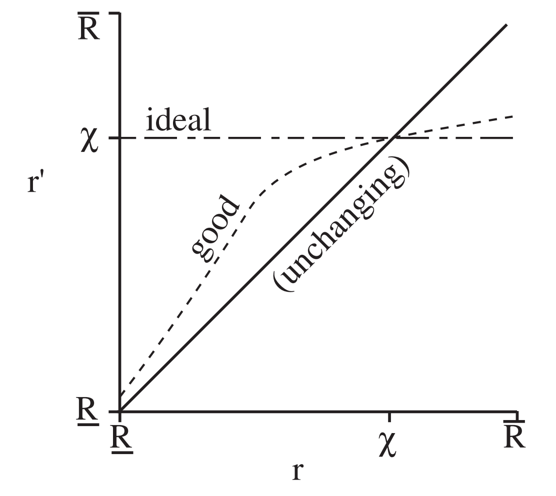
Trust System Metrics: Notation

- Agent type: $\theta \in \Theta$
- Current reputation (projection): $r \in [\underline{R}, \overline{R}]$
- Next reputation function: Ω
  - $r' = \Omega_\theta(r)$
- Fixed point reputation function: χ
  - $\chi(\theta) = \textsc{select}\{r \in [\underline{R}, \overline{R}] : r = \Omega_\theta(r)\}$
  - $\textsc{select}$ is max, min, second highest, etc. depending on Trust System
  - How to select $\textsc{select}$? ...

Dynamic Reputation Graphs



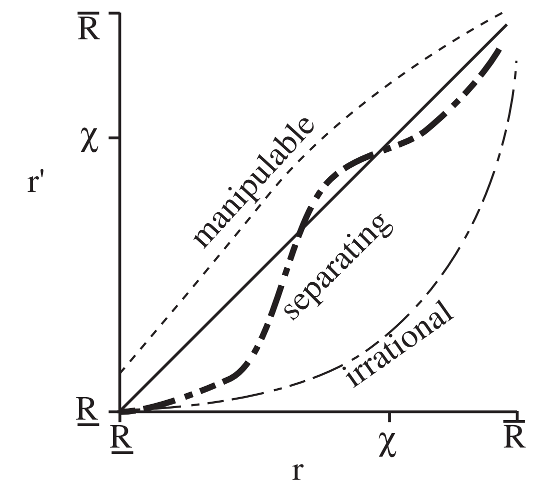
Ideal & Good Trust Systems



Trust System Metric 1: Unambiguity

- Each type should asymptotically map to a single reputation value
- $\forall \theta \in \Theta : |\{r \in [\underline{R}, \overline{R}] : r = \Omega_\theta(r)\}| = 1$
- If not, then reputation a combination of prejudice & meaningless

Ambiguous Trust Systems



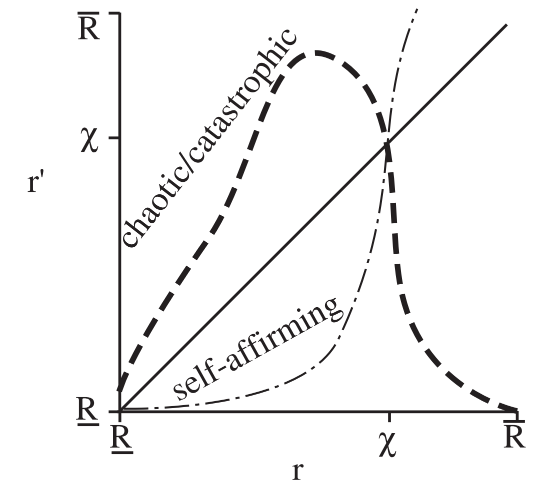
Trust System Metric 2: Monotonicity

- *Ideally Patient Strategic* (IPS) agent
  - Infinite horizon, maximize utility
  - IPS agent b, other agent a
  - $E(U_b(\theta_a)) = \lim_{\tau \to \infty} \max_{\sigma_b} \frac{1}{\tau} \sum_{t=0}^{\tau} u(t, \sigma_{b,t}, \theta_a)$
- If θ_a is weakly preferable to θ_b to IPS agent c, that is, $E(U_c(\theta_a)) \geq E(U_c(\theta_b))$, then a's asymptotic reputation should not be lower than b's reputation.

Trust System Metric 3: Convergence

- Reputation should converge quickly near the fixed point
- $\left|\frac{d\Omega}{dr}\right| < 1$ and minimized
- $\frac{d\Omega}{dr} < 0$: oscillate
- Lyupanov stability may be acceptable

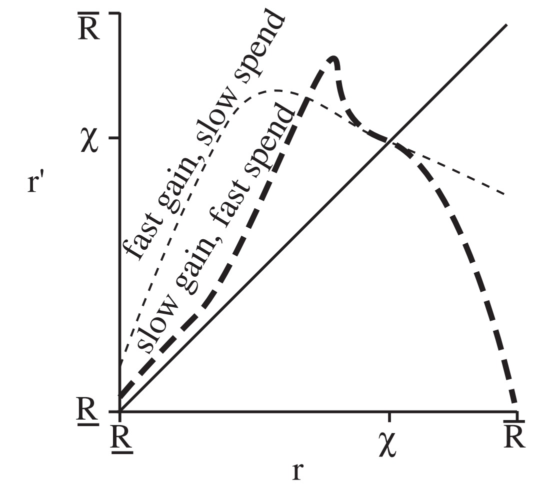
Non-converging Trust Systems



Trust System Metric 4: Accuracy

- Error: $\epsilon \in [0, 1]$
- Distance from ideal: $\epsilon_\theta(r) = \frac{|\chi(\theta) - \Omega_\theta(r)|}{\overline{R} - \underline{R}}$
- *Average Reputation Measurement Error* (ARME): $E(\epsilon_\theta) = \int_{\underline{R}}^{\overline{R}} \epsilon_\theta(r) dr$
- ARME minimized to distribution of types
  - PDF of θ, $f(\theta)$
  - minimize $E(\epsilon) = \int_\Theta f(\theta) \cdot E(\epsilon_\theta) d\theta$

Differing Accuracy



Performance Comparison

- Chose systems that
  - Measured reputation, not just aggregator
  - Diversity of models
  - Straightforward implementation
  - Connect reputation with decisions/utility
- Scenario
  - Take turns deciding to offer favors, one turn for each agent each round
  - Can spend own utility ($1-$12) to improve other's utility ($10-$30)
  - Agents discount the future (0.0 - 0.6)
  - Rational agents (moral hazard)

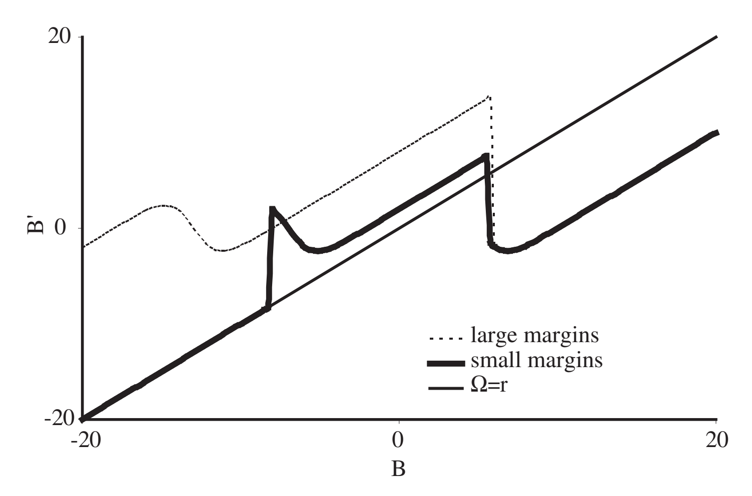
Utility & Decisions

- Probabalistic Reciprocity, Discount Factor: specify utility directly
- Others: utility based on reputation, per Zacharia & Maes '00
  - Linear relationship: risk neutral
  - sublinear relationship: risk averse
  - superlinear relationship: risk seeking

Probabalistic Reciprocity

- Sen '02
- Agent keeps ballance of favors
- Higher favor debt, lower cost of favor $\longrightarrow$ higher probability of offering favor
- Sigmoid function

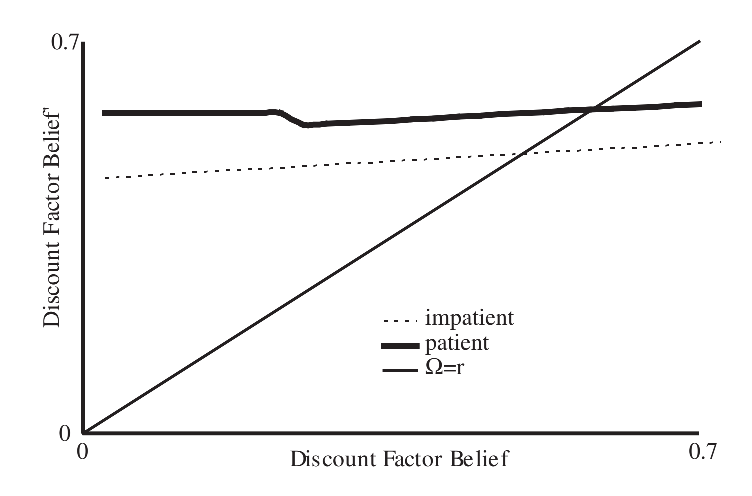
Probabalistic Reciprocity Graph



Discount Factor

- Hazard '08, Smith & desJardins '09
- Trustworthiness $\sim$ patience
- Model interaction from other agent's perspective based on future utility
- Assess constraints on discount factor (e.g. < 0.5)
- Use expected value of discount factor in modeling utility

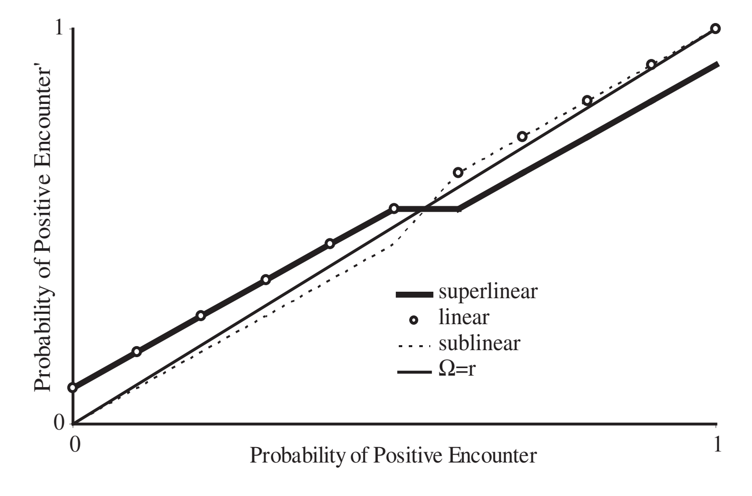
Discount Factor Graph



Beta Model

- Jøsang '98
- Quantize interactions into positive and negative
- Assume underlying probability agent will offer positive v negative result
- Model via Beta distribution

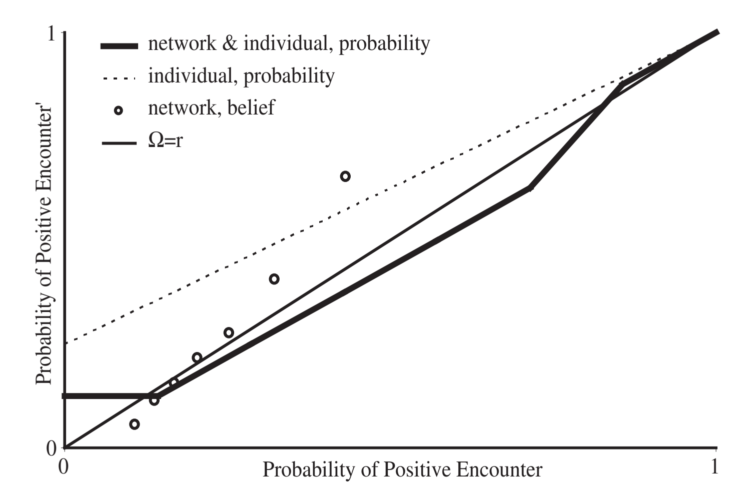
Beta Model Graph



Certainty Model

- Wang & Singh '06, '07
- Quantize to positive & negative like Beta model
- Use Dempster-Shafer model of evidence-based belief: probability & uncertainty
- Also tested against group of 3 agents, aggregating evidence

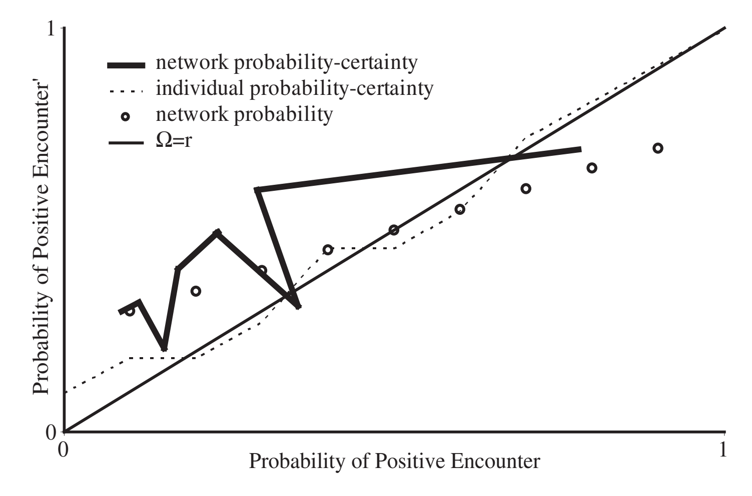
Certainty Model Graph



TRAVOS Model

- Teacy, Patel, Jennings, Luck '06
- Quantize to positive & negative like Beta model
- Subdivide reputation space into 5 regions (Beta distribution), find region with largest area under PDF, largest area is certainty
- To communicate reputation, normalize magnitude preserving mean and standard deviation

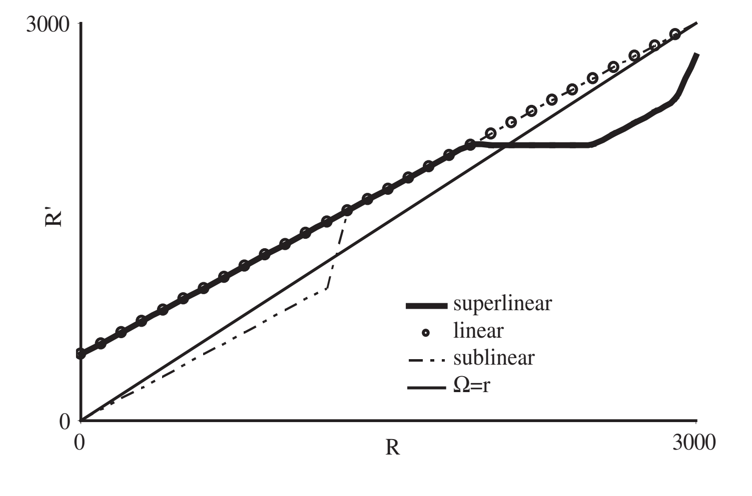
TRAVOS Model Graph



Sporas Model

- Zacharia & Maes '00
- Reputation measured on range
- Ratings dampened with new measurements

Sporas Model Graph



Results

Trust System	Unambig.	Monotonic	Converge	Accuracy
Prob. Reciprocity	no	yes	no	0.2
Discount Factor	yes	yes	< 0.1	0.02
Beta	no	no	no+	.3
Certainty	weakly*	yes	0.9	0.37
TRAVOS	no	yes	0.9	0.32
Sporas	no	no	no	0.31

*weakly unambigous means ambiguous points difficult to reach

+converged on superlinear case

Conclusions

- Trust system metrics useful for comparison within a domain
- Discount Factor shows considerable promise, but does not yet support non-discrete choices
- Desiderata and metrics presented are not the final word
  - Are IPS agents the best comparison for monotonicity?
  - Absolute mean deviation best error measure?
  - Evaluating multi-context models

On Computational Trust... (2)

- "Never trust anything that can think for itself if you can't see where it keeps its brain." - J.K. Rowling, *Harry Potter and the Chamber of Secrets*