Modeling and Learning Contextualized Probabilistic Trust

Matthias Nickles and Achim Rettinger

Distributed Artificial Intelligence Group
Department of Computer Science
Technical University of Munich

Introduction

- Autonomous agents in *social interaction systems* which are o*pen, heterogeneous* and *complex*
- Examples: Open multiagent systems and hybrid interaction systems (where humans collaborate with artificial agents), such as trading platforms on the Web
- We take the viewpoint of one of the agents in such a system
- Agent observes its interaction environment and tries to compute an optimal action policy
- Talk focuses on the learning aspect, not planning or reasoning
- We are particularly interested in the machine learning of so-called *initial trust…*

Introduction

- A few challenges:
  - Peer agents may show a highly contingent behavior
  - Incomplete, uncertain or conflicting information from peer agents
  - Unknown semantics of observed interaction data
  - Interaction with previously unknown agents
- These issues originate from…
  - Agent autonomy. Agents acts under self-control towards hidden goals, with hidden beliefs and intentions
  - Absence of reliable rules and other prior knowledge
  - Possibly: partial observability of environment, noise

Introduction

- In the described systems, trust is crucial in information exchange, negotiation, and any other kind of social interaction
- To trust (or not to trust) is seen as a crucial mechanism in order to allow informed *decision making*
- From the viewpoint of a rational learning agent, trust is essentially a means to make the behavior of peer agents *predictable*
- More precisely: trust, as we understand it in this talk, can be boiled down more or less to the predictability of the fulfillment of some given promise (commitment)

Introduction

- Trust can be approached, e.g., using
  - Rules and regulations enforced by the system
  - Privacy mechanisms: encryption and certificates
  - Reputation and recommendations
  - Incentives of the opponent
  - Own observations of the opponents behavior
- Many of these approaches (try to) enforce or ensure trustability normatively
- In contrast, our approach is descriptive - it focuses on the prediction of agent and human behavior
- How to obtain trust-related predictions? (Or how to obtain predictions which indicate distrust, respectively?)

Contextualized interaction trust

- We look mainly at so-called *initial-trust situations*
- Absence of trust- or fairness-ensuring mechanisms, enforceable (hard) norms…
- Especially in open systems, the agent possibly needs to interact with more or less unknown (new) agents in an ad-hoc manner
- No or not much information available which relates directly to trustability, such as…
  - transitive trust (e.g., in a trust network)
  - past interactions of the same kind, or with the same agent
  - reputation
- Property of such situations: the agent does <u>not</u> have sufficient information that could directly be applied to assess trustability

Contextualized interaction trust

- That is, the agent needs to learn trust entirely from short term observations of previously unknown peer agents and the context of the respective interaction
- It needs to do so rapidly
- Learning features do not directly relate to trust
- Observed interaction history at hand does not necessarily relate directly to the respective trust situation
- Furthermore, trust is typically *multi-dimensional*…

- These problems have not been sufficiently addressed by traditional machine learning approaches to trust

Contextualized interaction trust

- Basic idea: Infer trust (distrust, respectively) more or less ad hoc from the entire available interaction context
- Let the machine learning algorithm sort out automatically which contextual aspects are important for trust learning, and which are not
- More concretely:
  - Infer trust from all available contextual information, such as [dependencies between] attributes of opponents and the interaction situation
  - *Transfer knowledge* from previous indirectly related contexts (e.g., interactions with other agents) to new situations
- Context is thus not seen as yet another parameter of trust (as usual), but as crucial learning input

Contextualized interaction trust

- Technical approach:

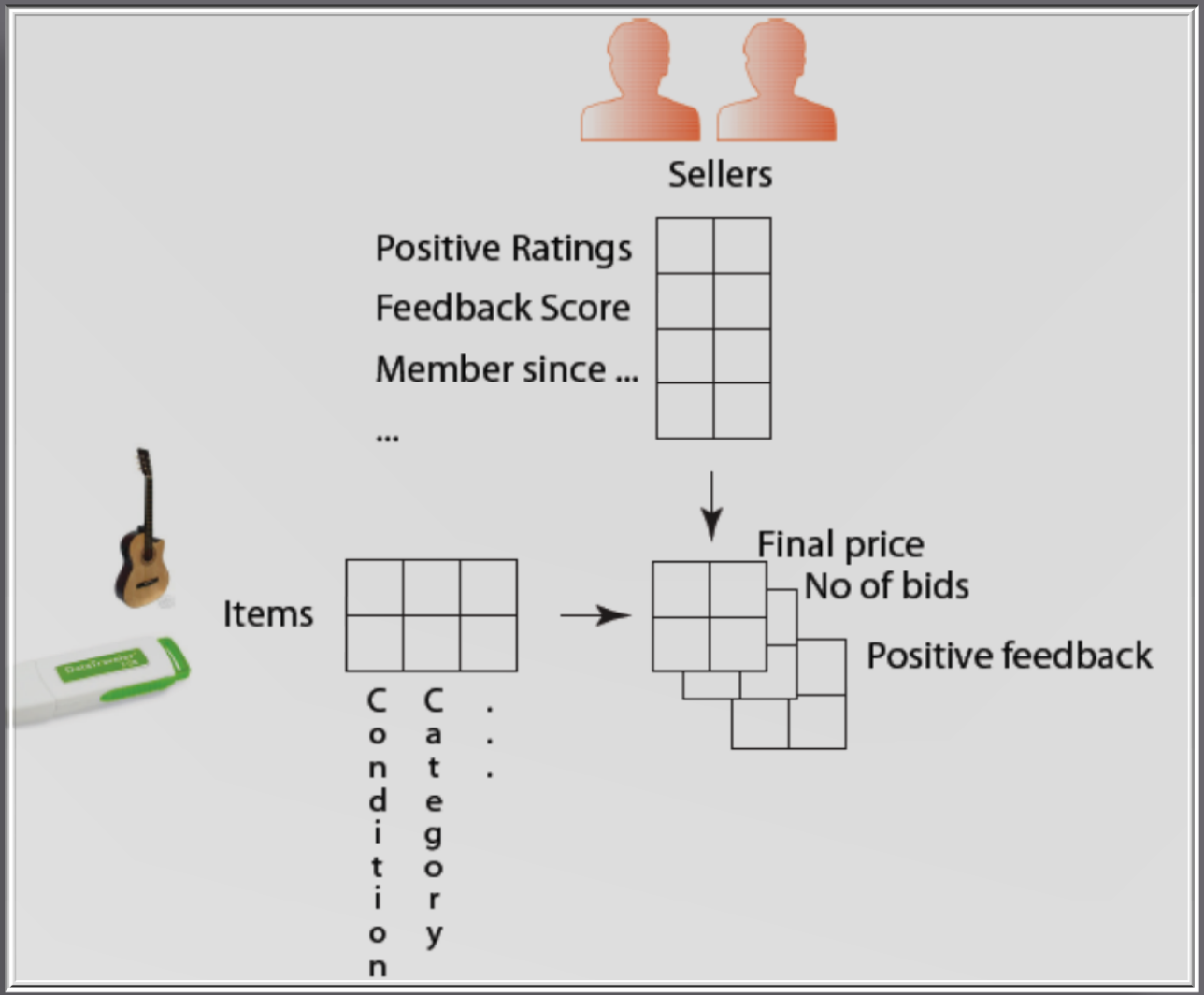
  Infinite Hidden Relational Trust Models (IHRTMs)
- Based on *Infinite Hidden Relational Models* (IHRM), a recent and very powerful approach to *statistical relational learning*
- Takes into account features which relate directly to trust (such as reputation, if available), but - even more importantly - also other contextual features
- Takes into account personal features, but also the social interaction itself
- Takes into account any number of trust dimensions

Machine learning of contextualized interaction trust

- Modeled scenario:
  - An agent *a* (trustee), characterized by a set of observable attributes Att^a . E.g., a seller on eBay and its properties
  - A set of conditions of the environmental state *s* with corresponding attributes Att^s. E.g., the items for sale from the trustee
  - A relation *interacts(a,s)* with a set of relationship attributes Att^c and Att^t.
    - Att^c describe the mutual commitments of the trustee and the truster, such as the agreed price for the merchandize
    - Trust attributes Att^t could include any dimension(s) of trust, e.g., whether after a merchandize the bought item is as described by the seller, or the seller rating (feedback)
      These attributes are essentially predicted by our algorithm.

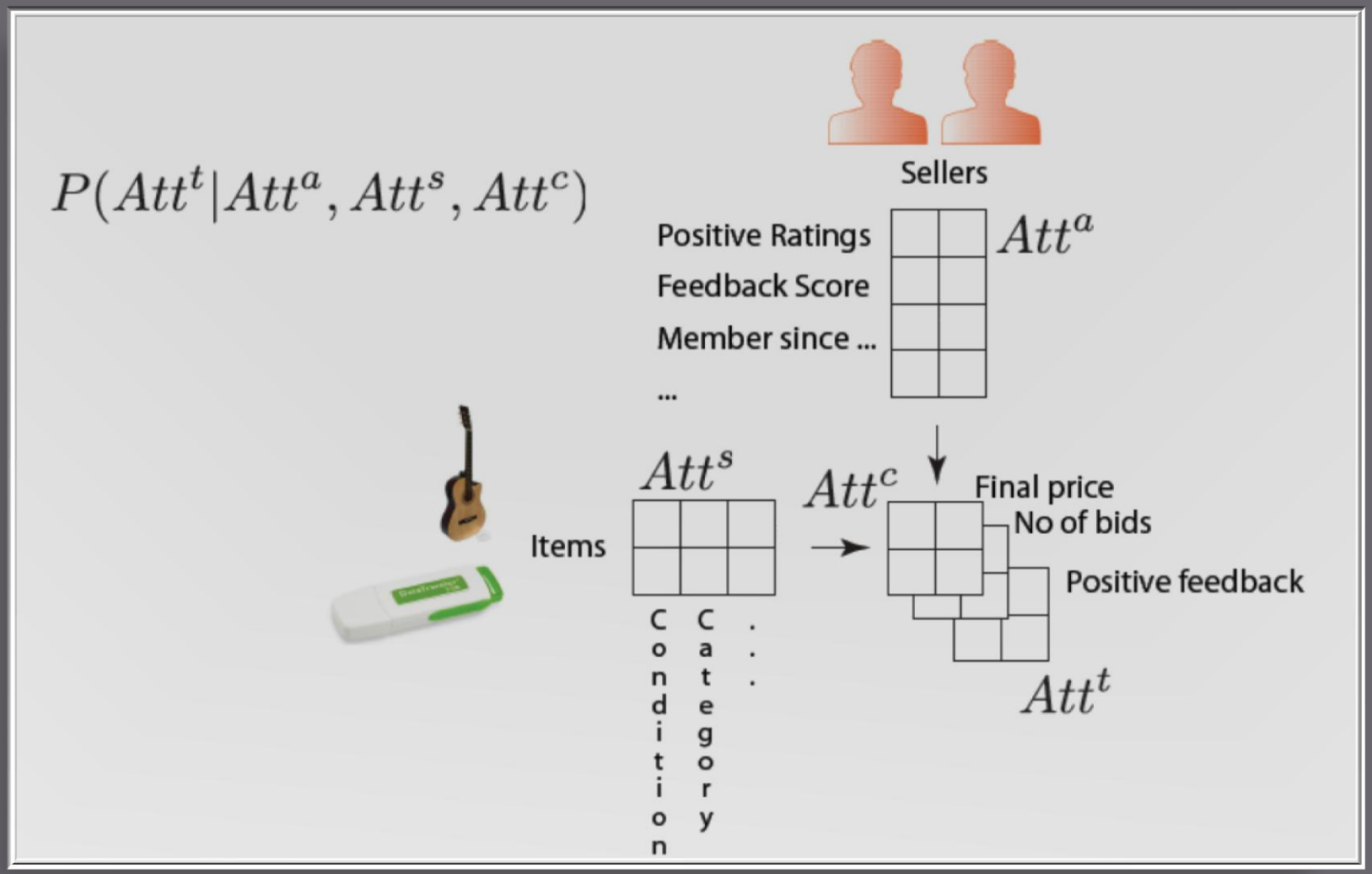
Contextualized interaction trust

- Scenario overview:

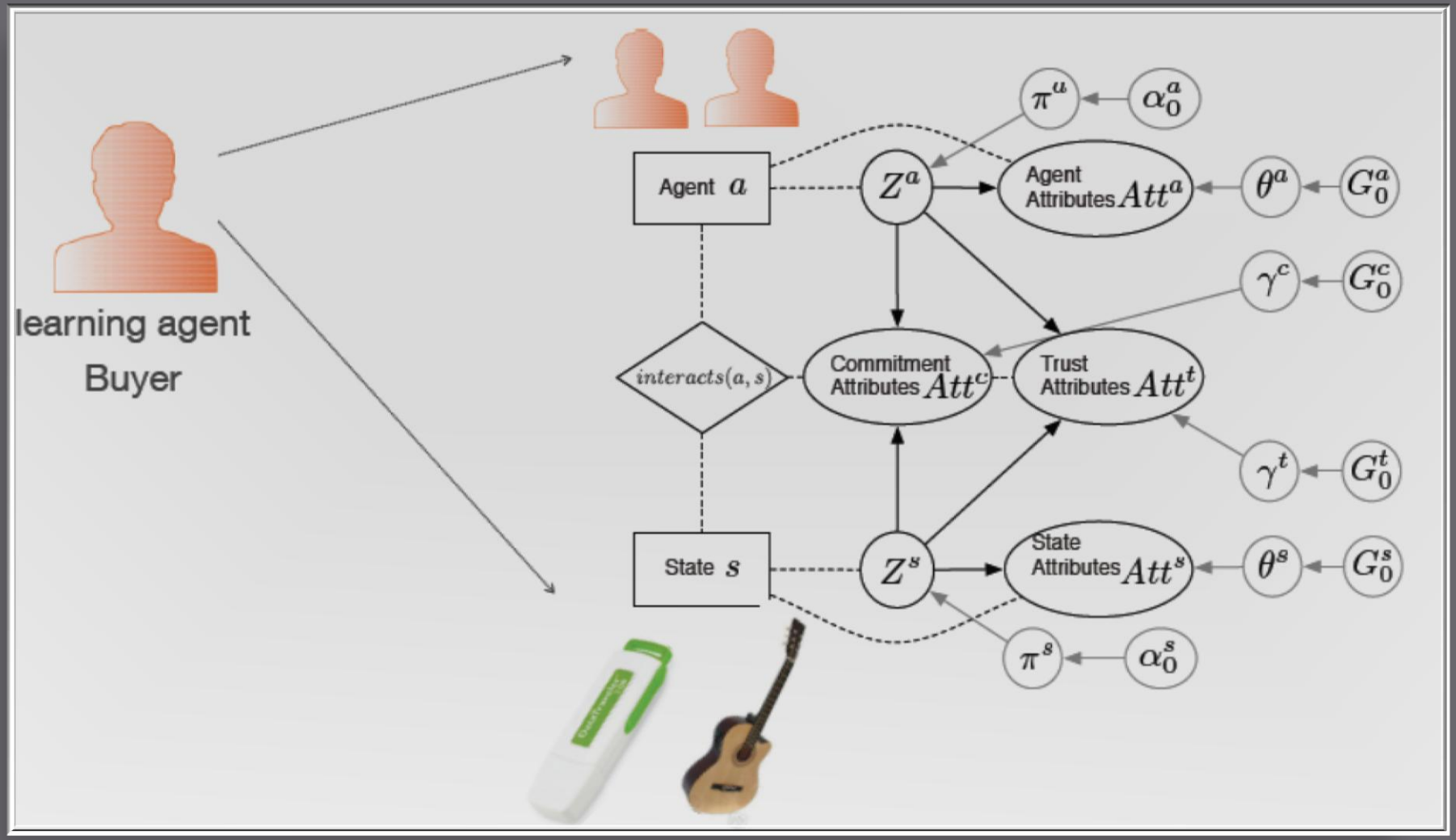


Contextualized interaction trust

- Scenario overview:



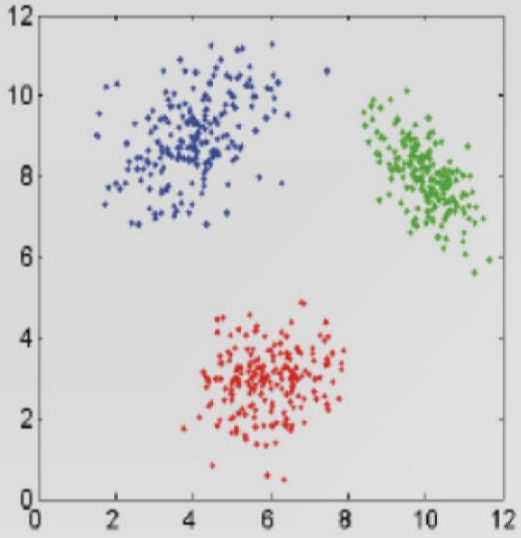
Machine learning of contextualized interaction trust



Machine learning of contextualized interaction trust

Non-parametric Bayesian mixture model

- Mixture model:
  - Data was generated from a mixture of several, distinct distributions
  - Can be interpreted as a soft clustering
- Latent/Hidden variables:
  - Make Clustering explizit: Z
  - Multinomial latent (unobservable) random variable
  - Prior: Dirichlet prior
- Infinite variables:
  - Possibly infinite number of clusters
  - Prior: Dirichlet process



Machine learning of contextualized interaction trust

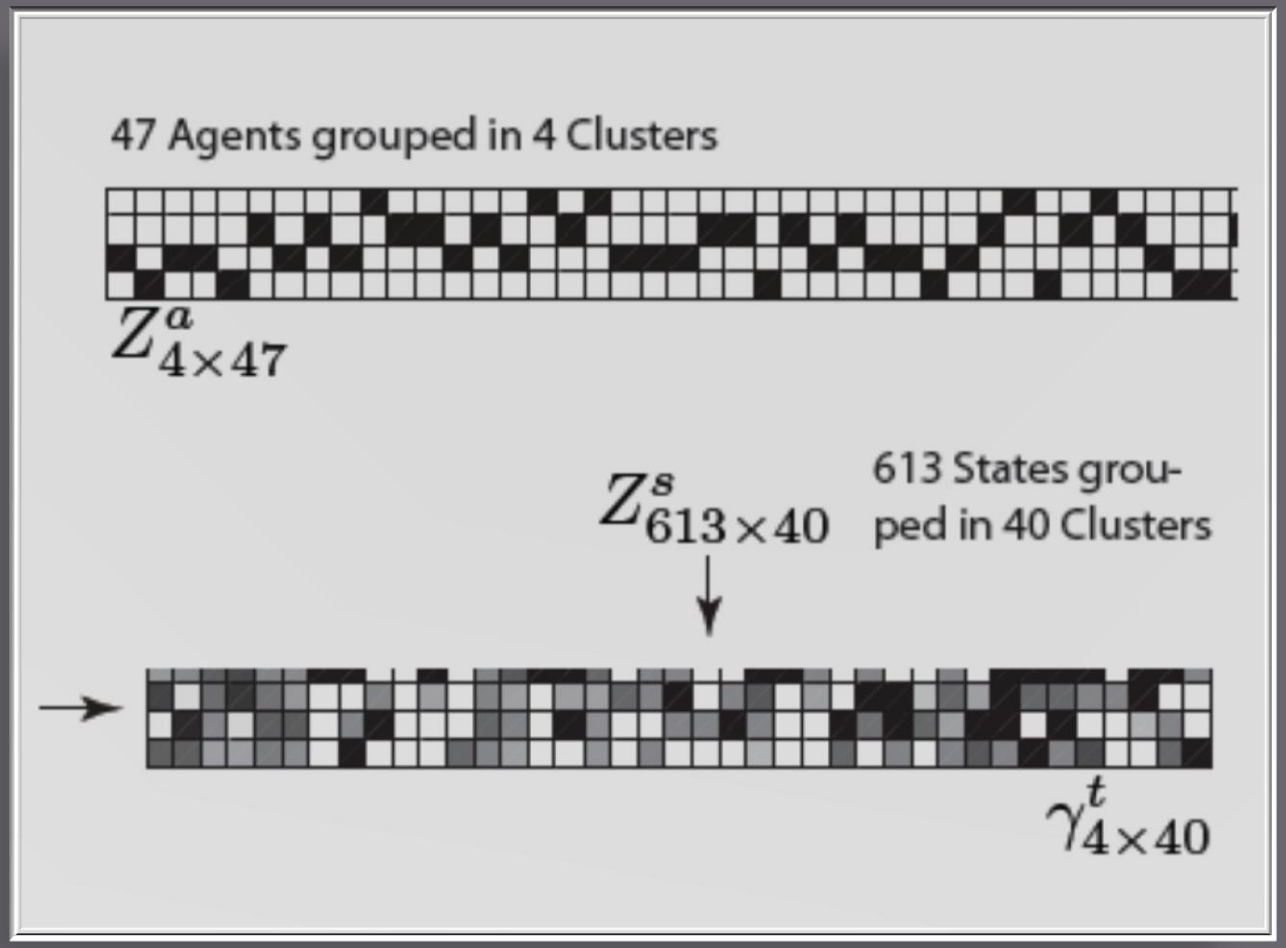
- Technical properties of an IHRTM
  - Goal is to group entities into clusters
  - IHRTM automatically discovers number of clusters and cluster assignments - at runtime
  - Cluster assessment is influenced by both attributes and existing cluster assignments
  - Cross-attribute and cross-entity dependencies can be learned automatically

Machine learning of contextualized interaction trust

- Experiments (1):
  - 47 sellers (with at least 10 negative ratings)
  - 630 items (from 47 categories)
  - 1818 rated sales

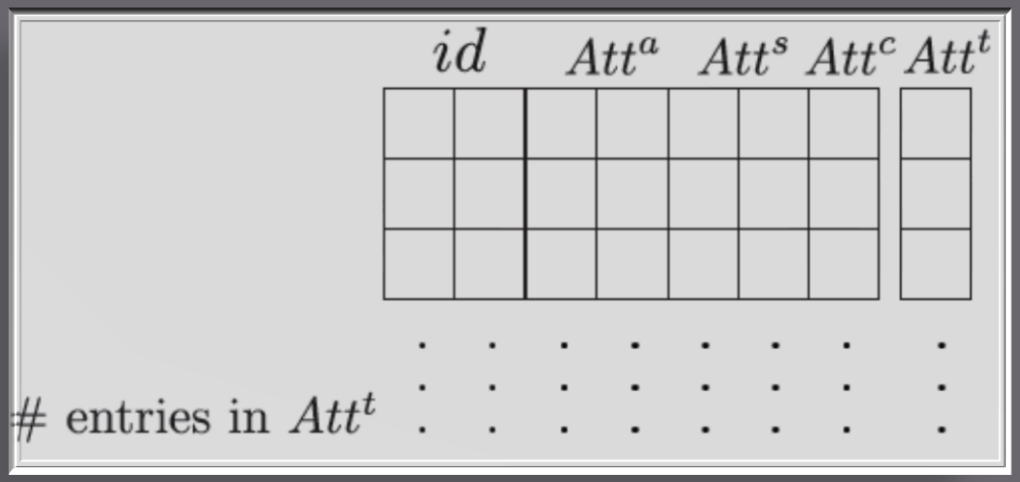
Machine learning of contextualized interaction trust

- Clustering of trustees (sellers on eBay) and derived probabilities of being trustworthy:



Machine learning of contextualized interaction trust

- Comparison to propositional machine learning algorithms (SVM, Decision Tree)
- Propositionalization needed:
  - all attributes (+ label)
  - all attributes + id (+ label)



Machine learning of contextualized interaction trust

- Predictive performance of IHRTM:

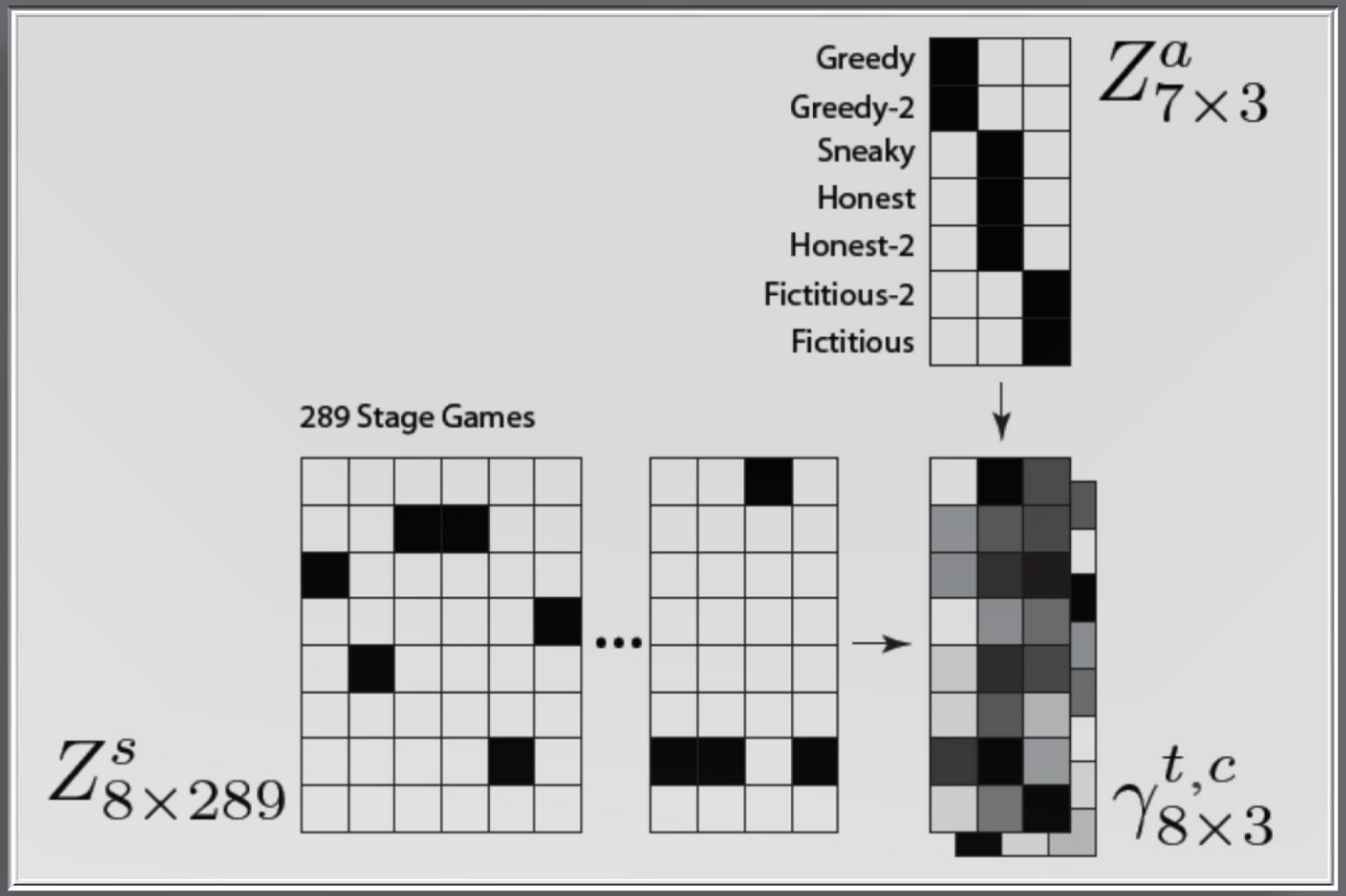
	Accuracy
Ratio	48.5334 (±3.2407)
SVM	54.1689 (±3.5047)
DecTree	54.6804 (±5.3826)
SVM+ID	56.1998 (±3.5671)
DecTree+ID	60.7901 (±4.9936)
IHRTM	**71.4196** (±5.5063)

Machine learning of contextualized interaction trust

- Experiments (2) - Negotiation game:
  - 2 agents trade resources to optimize collection of goods
  - 3 phases: Negotiation - Trading - Evaluation
  - Players:
    - 3 agent types (Honest, Greedy, Fictitious)
    - 2 negotiation strategies (monotonic concession protocol)
    - 3 different trading strategies (Fictitious is non-stationary)
  - Data samples:
    - 600 observed interactions (200 per agent type)
    - 289 different repeated games (available resources before negotiation-phase)

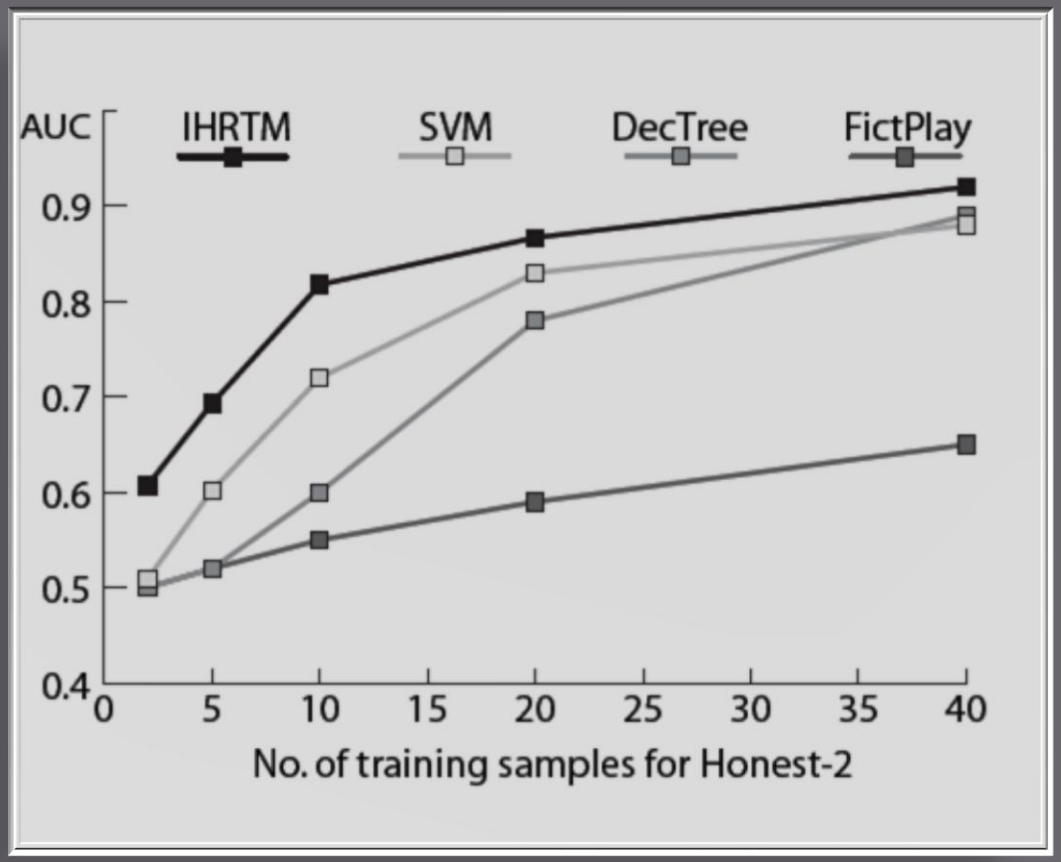
Machine learning of contextualized interaction trust

- Clustering:



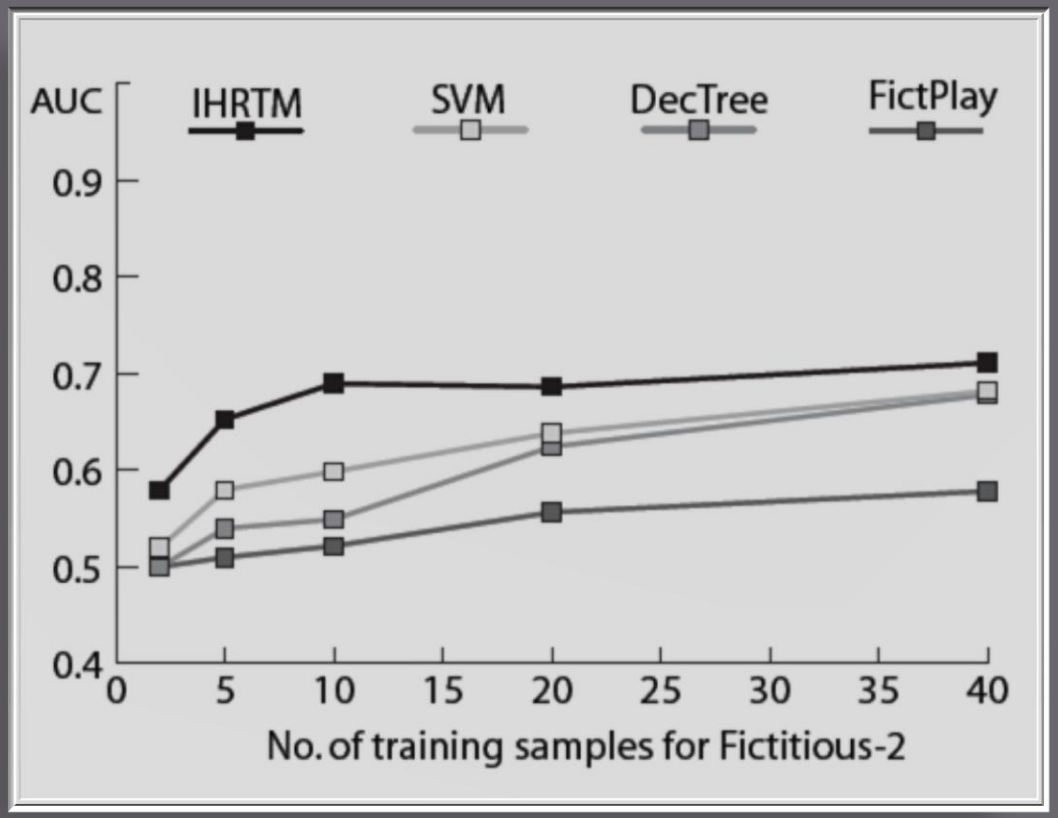
Machine learning of contextualized interaction trust

- Predictive performance:



Machine learning of contextualized interaction trust

- Learning efficiency:



Conclusions

- Features of Statistical Relational Trust Modeling
  - Trust learning in initial trust situations where no direct trust information (such as reputation) is available
  - Entire context of the respective interaction is used for the learning tasks
  - Meaningful trust assessment by inherent clustering and improved predictive performance by collaborative filtering
  - Rapid adaptation and fast learning using knowledge transfer. Improved "cold start" performance
- Future work:
  - Further evaluations
  - Integration with logical frameworks and additional (a-priory) knowledge and heuristics about trust

Thank you very much for your attention.