

Reasoning about Rationality and Beliefs

Ya'akov Gal
Avi Pfeffer

Game Theory and MAS

Game Theory assumptions	Real world agents
- agents are rational - common knowledge of game structure - agents' beliefs are correct/ consistent	- agents play irrationally - agents are uncertain about game, others' strategies - agents' beliefs might be incorrect

Bayesian Games

- Represent uncertainty over decision-making models with types.
- Contrived, unnatural and large representation
 - Must represent a joint distribution over all type instantiations.
 - Assume agents are rational.
 - Cannot naturally model heuristics.
 - Number of types might be exponential in number of players.
 - Real-world variables folded into utility functions.

Network of Influence Diagrams (NID)

- A language for reasoning about uncertainty over decision-making processes.
- Distinguishes between the real world and mental models used to deliberate.
- Original formalism [Gal and Pfeffer *AAMAS03*]
 - Could express a subclass of Bayesian games.
 - Provided algorithms for computing best-response.
 - Assumed agents' beliefs are correct.

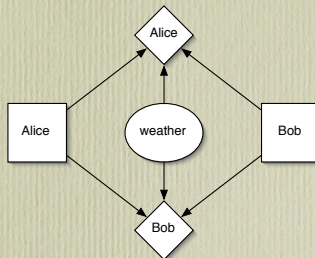
Contributions

- Enhanced expressivity of NIDs; can model
 - conflicting beliefs
 - irrational behavior
 - cyclic belief structures
- Formalized NID equilibrium
- Relationship with Bayesian games

Battlefront Scenario

- Generals Alice and Bob are getting ready for battle.
- Either general can instruct a surprise attack, or she can defend.
- General Alice will likely win if both/none armies attack.
- General Bob will likely win if one army attacks and the other defends.
- Success of attack depends on weather conditions.

Multi Agent Influence Diagrams (MAID)



- MAIDs [Milch, Koller '00] describe multi-agent interactions.
- Graph topology describes structure of game.
- Solving MAID – computing Nash equilibrium

Example extended

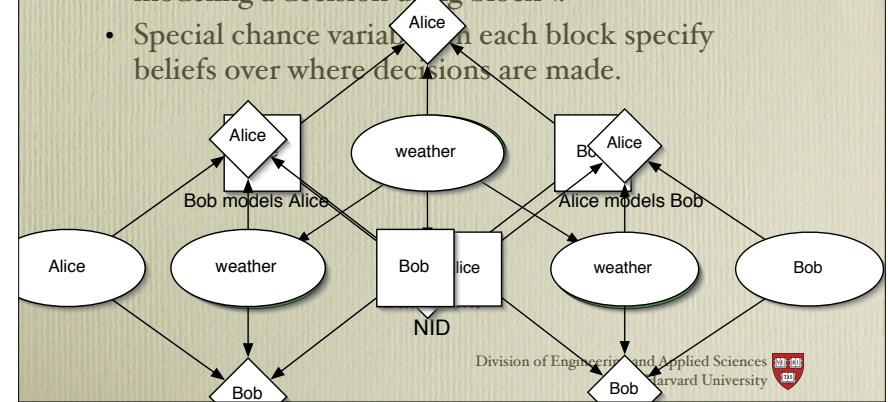
- Experts advise Alice to attack and Bob to defend.
- With probability 0.7, Alice believes Bob will follow experts.
- With probability 0.3, Alice believes Bob will play best response to his model.
- Alice's beliefs are symmetric.
- To capture this, we need to express uncertainty over decision making models.

Network of Influence Diagrams

- A NID is a set of mental models describing a game.
- Models may differ in
 - Utility functions
 - Distributions over random variables
 - Decisions that are replaced by automaton.
- Each mental model is represented by a MAID.
- Beliefs over mental models are directly represented.

NIDs graphical representation

- Edge (u,v) implies some agent in block u is modeling a decision using block v .
- Special chance variables in each block specify beliefs over where decisions are made.

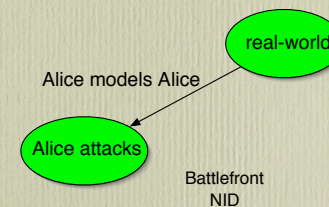


Modeling irrational behavior

- Suppose there are 2 successive battles.
- Each general is faced with a sequential decision problem.
- Alice believes with probability 0.3 to be affected by political pressure and attack in the second battle.
- Important for Alice to reason about her irrational behavior in the second battle.

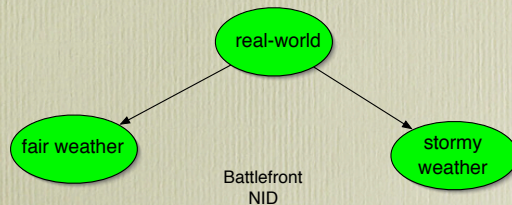
Modeling irrational behavior

- NID captures agents who play differently than best-response.



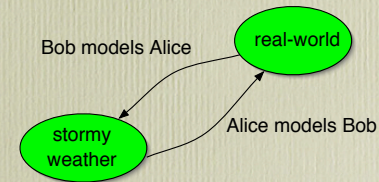
Modeling conflicting beliefs

- Weather affects chances of success for battle.
- Alice believes the weather to be fair with high probability. Bob believes the weather to be stormy with high probability.



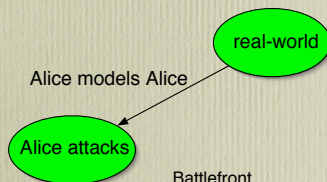
Modeling cyclic beliefs (I believe that you believe that I believe...)

- With prob. 20%, Bob believes that Alice believes that likelihood of stormy weather is high.
- Bob believes that Alice reasons about Bob modeling her.



General NID equilibrium

- Equilibrium includes two types of strategies:
 - Actually-played – model's prediction of how agents actually behave.
 - best-response - strategy that maximizes outcome given agent's model.
- Actually-played and best-response strategies defined in terms of each other.
- Equilibrium expressed as set of fixed point equations.



NID equilibrium algorithms

- Find best-response and actually-played strategies at each model in the NID.
- In general, difficult to find solution in closed form.
- For acyclic NIDs, can solve using a bottom-up solution algorithm.



NIDs and Bayesian games

- Are NIDs and Bayesian games equivalent?
- Can we map (mental model, agent) in NID to a type in BG ?
- Every Bayesian game is a NID.
- Some NIDs are Bayesian games.
- Equivalent Bayesian game is contrived and large.



Current and future work

- NIDs can be used for opponent modeling in repeated games
 - Rock-paper-scissors competition [Gal and Pfeffer 03]
 - Automated negotiation with people [Gal et al 04]
- Temporal NIDs.
 - Representation, inference and learning.

Conclusions

- NIDs are a language for representing uncertainty over decision-making strategies.
 - Might be exponentially smaller than traditional representations.
 - Allow querying the network.
 - Can model rational and irrational agents.
 - Can model conflicting beliefs.
 - Can model cyclic belief structures.

Ya'akov Gal and Avi Pfeffer
{gal,avi}@eecs.harvard.edu



Thanks to the Harvard AI research group !