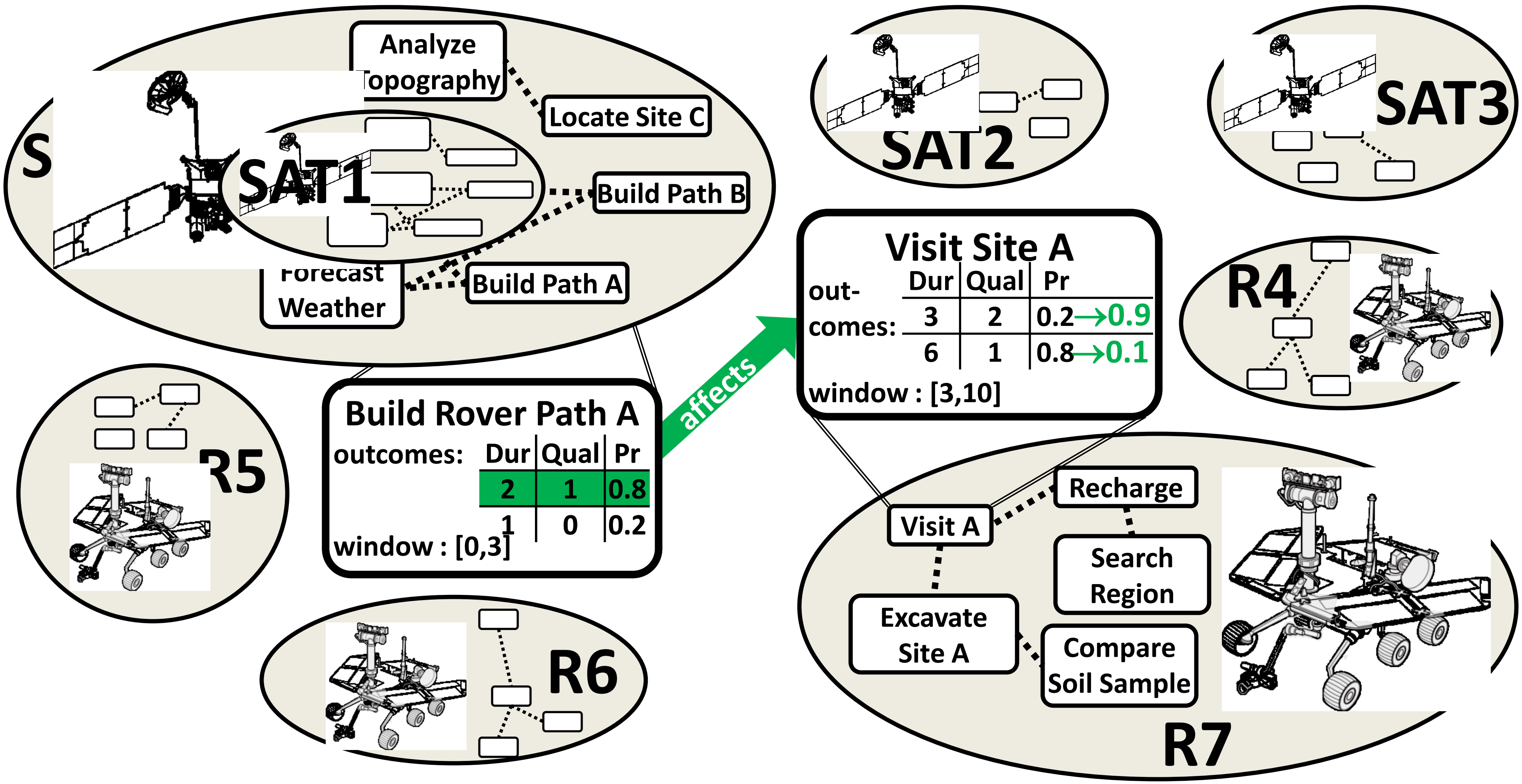
Influence-Based Policy Abstraction
for Weakly-Coupled Dec-POMDPs

Stefan Witwicki
witwicki@umich.edu
Ed Durfee
durfee@umich.edu

UNIVERSITY OF MICHIGAN
university of michigan artificial intelligence laboratory

Team Coordination Under Uncertainty

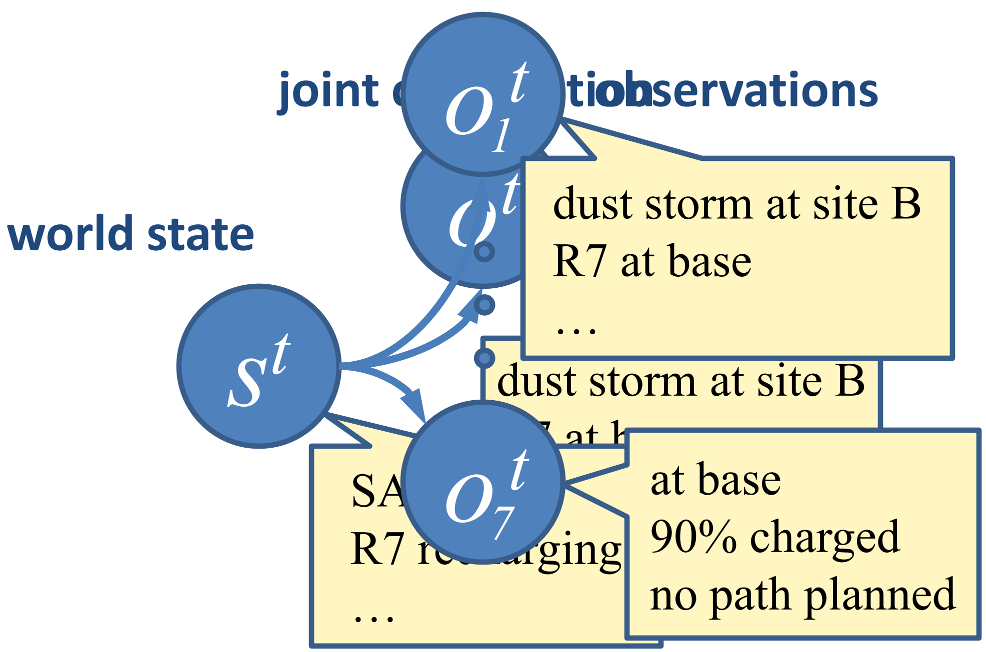
- System composed of weakly-coupled agent-controlled components
- Problem: plan agents' behavior so as to accomplish team objectives



Dec-POMDP

(***D**ecentralized **P**artially-**O**bservable **M**arkov **D**ecision **P**rocess*)

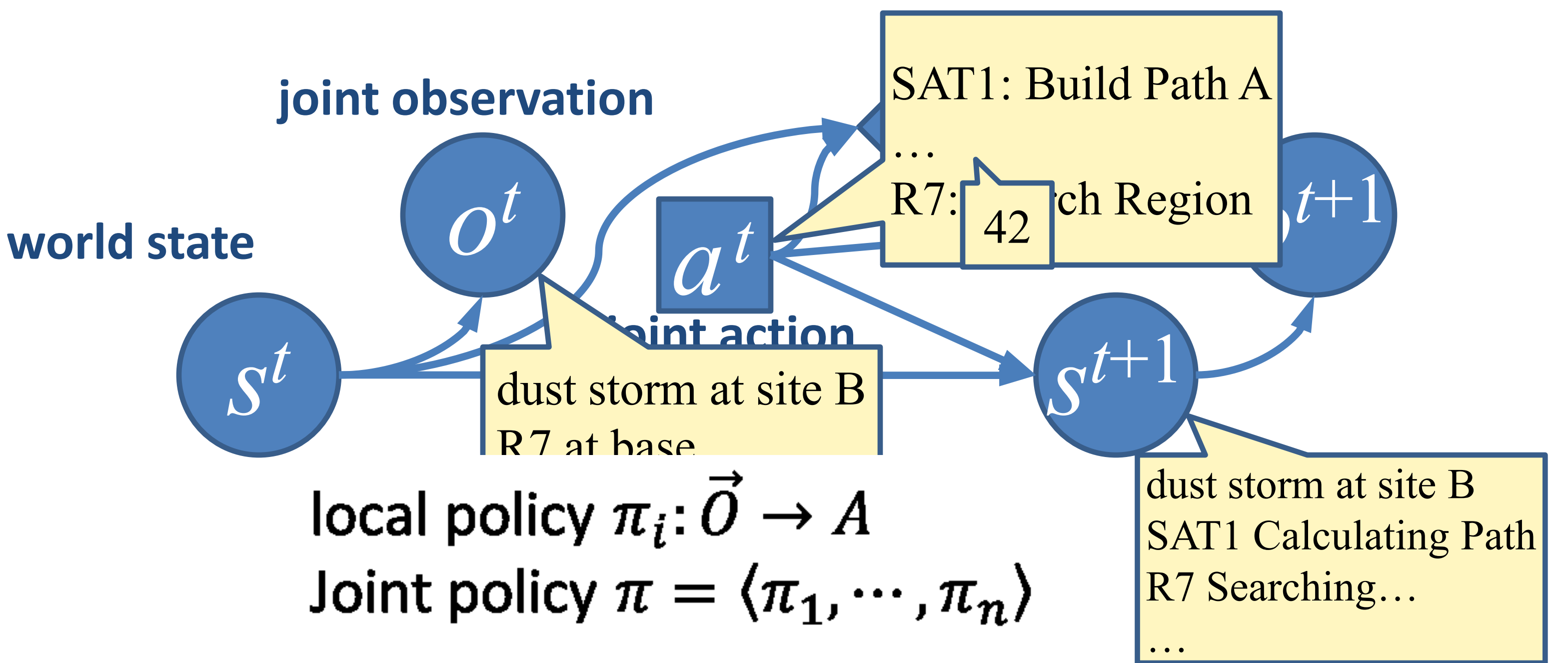
- Dec-POMDP is theoretically-appealing model for team coordination
  - decentralized / partial observations



Dec-POMDP

(***D****ecentralized* ***P****artially-***O****bservable* ***M****arkov* ***D****ecision* ***P****rocess*)

- Dec-POMDP is theoretically-appealing model for team coordination
  - decentralized / partial observations
  - outcome uncertainty
  - general, well-defined notion of optimality (reward model)



local policy $\pi_i : \vec{O} \rightarrow A$

Joint policy $\pi = \langle \pi_1, \cdots, \pi_n \rangle$

Motivation

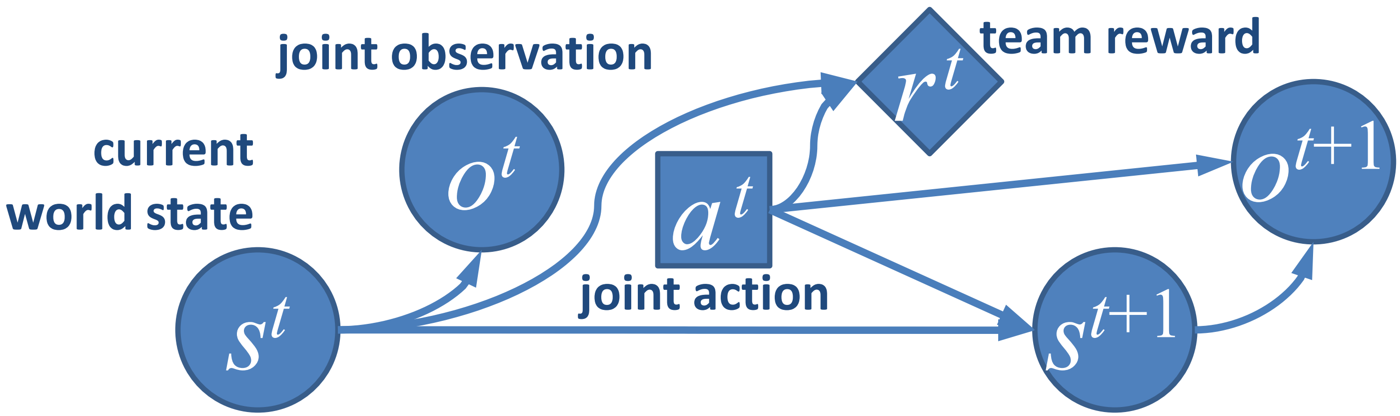
- Dec-POMDP is theoretically-appealing model

 …but <u>very</u> challenging to solve!
  - In general, NEXP ($\supseteq$NP, $\neq$P) complete $\Rightarrow$ **intractable**
  - State-of-the-art solution methods have not scaled beyond 3 agents, except by…
    1. Disallowing agent *interaction* through the transition and observation model (e.g. TI-DEC-MDPs [Becker *et al*], ND-POMDPs [Nair *et al*, Varakantham *et al*, Kumar *et al*])
    2. Restricting agents' *local* behavior (e.g. OC-DEC-MDPs [Beynier *et al*, Marecki *et al*])
    3. or Giving up on optimality and near-optimality (e.g. TREMOR [Varakantham *et al*])
- *Can we increase quality-bounded agent scalability while still allowing some general form of transition dependence?*

Our Contributions

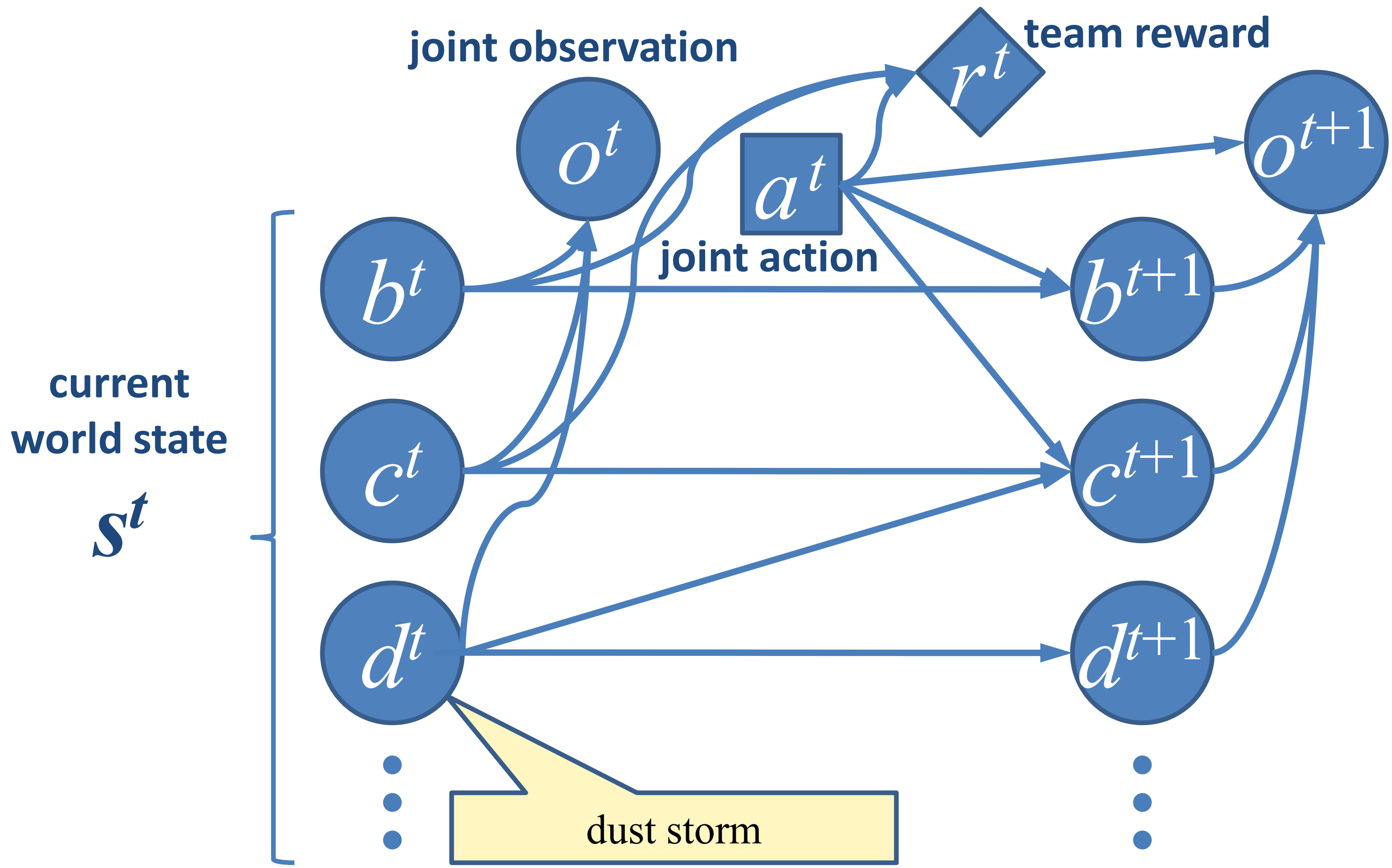
- Identification of exploitable transition-dependent interaction structure
- Characterization of abstract transition influences
- Algorithm for planning/coordinating optimal influences
- Empirical comparison with state-of-the art policy search methods

Dec-POMDP Model



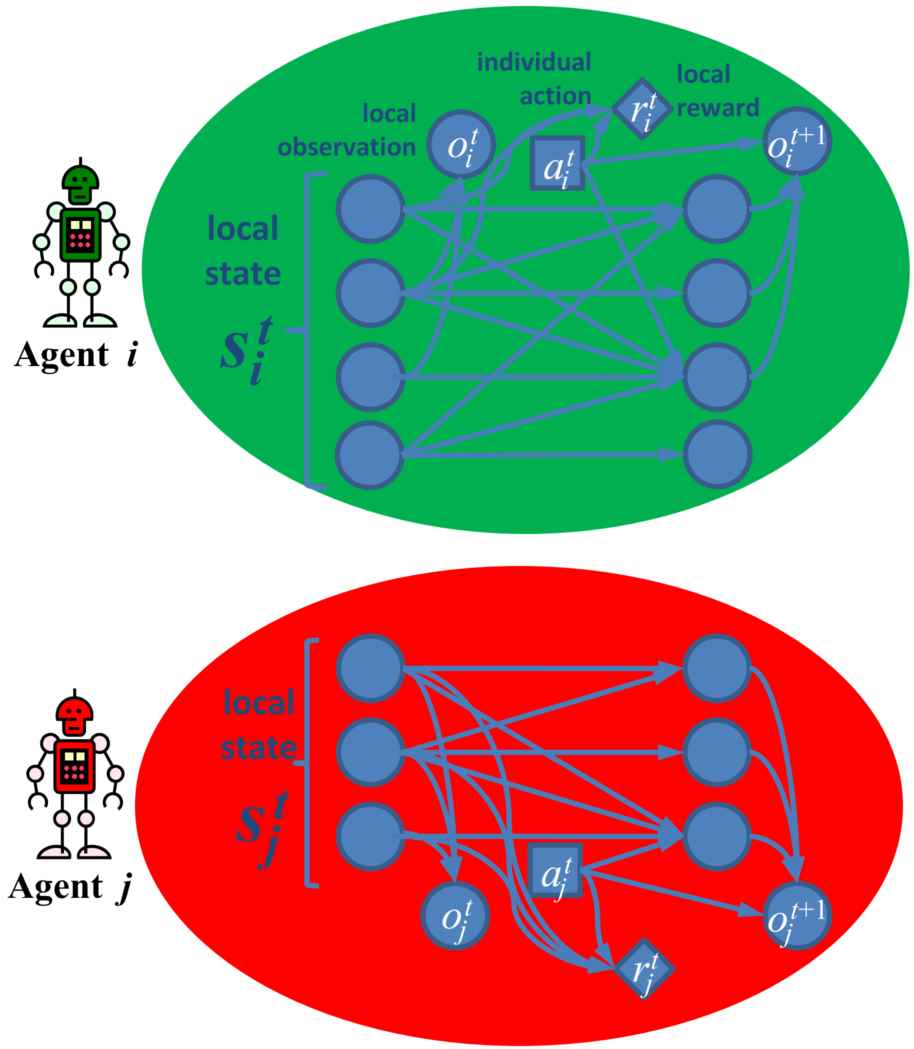
2-stage (Object-Oriented) Dynamic Bayesian Network

Factored Dec-POMDP



Extreme Factoring

- Imagine **fully-independent** agents, each modeling the world with a single-agent POMDP…
  - → world state is factored into **local state** feature subsets
  - → transitions are factored, and independent
  - → joint observations are factored, and independent
  - → team reward is factored into local rewards



TD-POMDP model

(***T**ransition **D**ecoupled POMDP*)

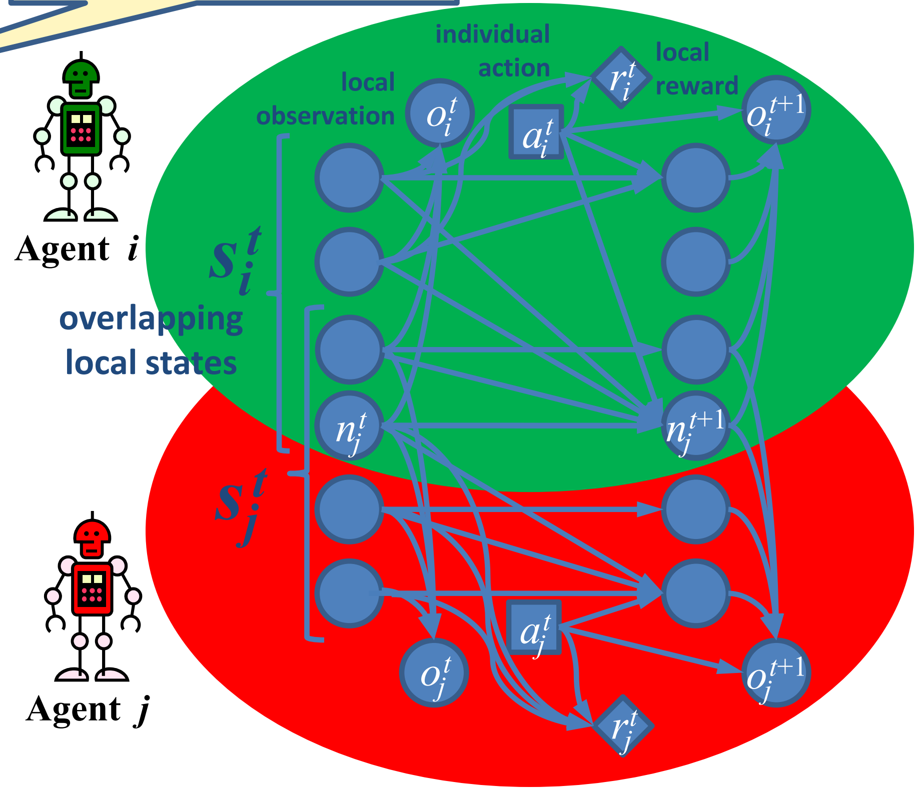
- Explicitly represent interaction

  via shared features…

  e.g. *Path-A-Built*

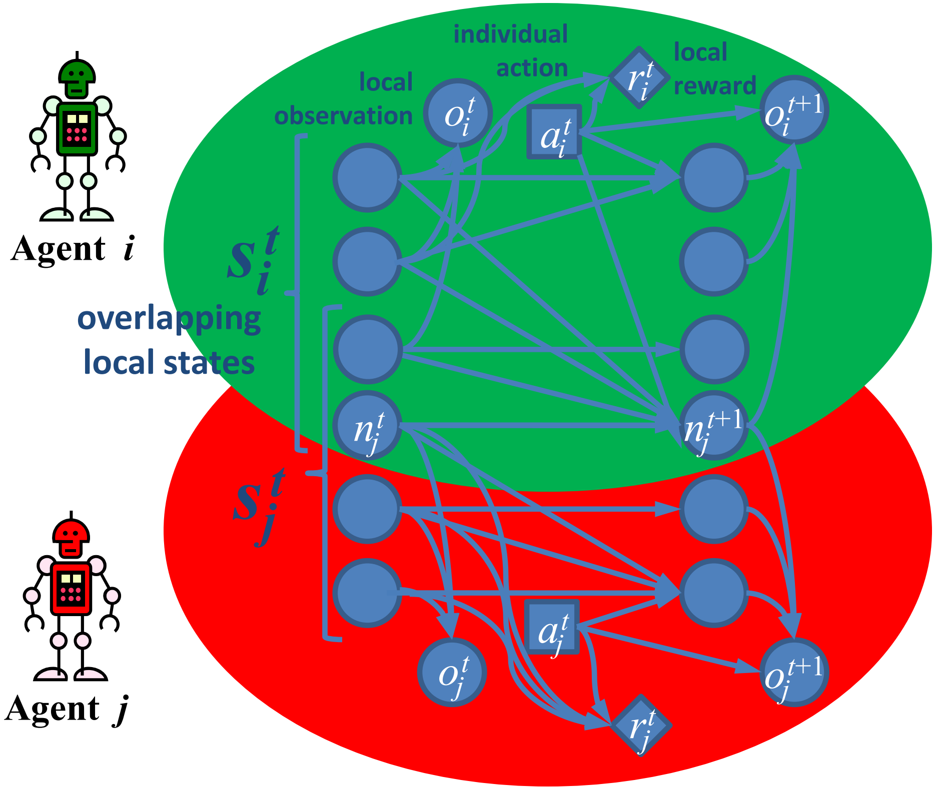
  - nonlocal feature n_j
    - controlled by another agent
    - affects subsequent transitions of other features in agent *j*'s local State

→ Agents are "transition-dependent", as well as "observation-dependent"



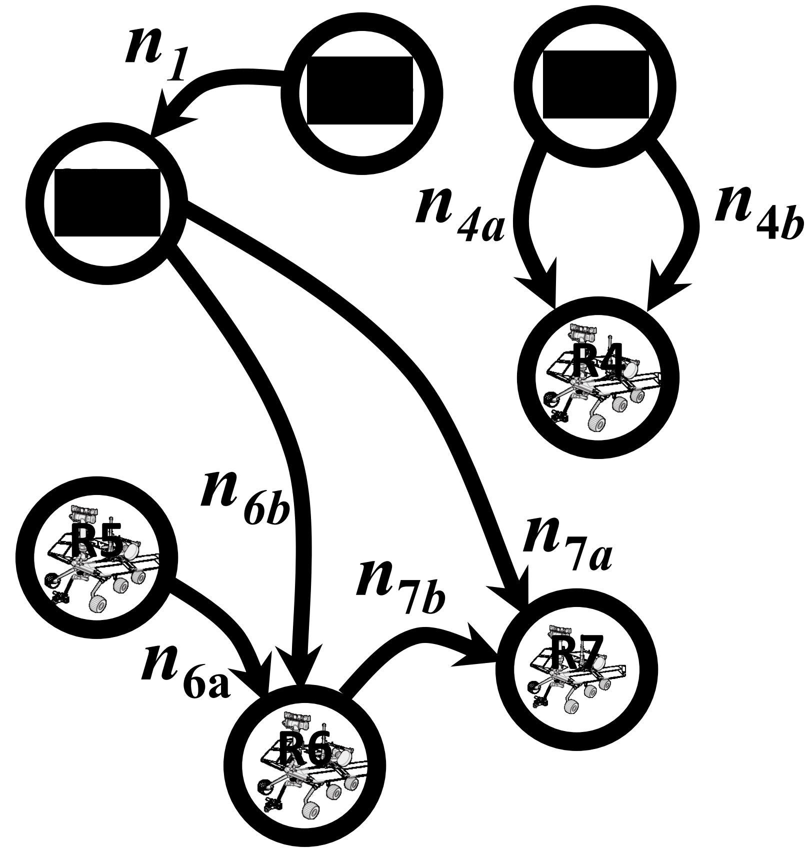
TD-POMDP Benefits

- Explicit representation of transition-dependent interaction features
- Naturally conveys
  - locality of interaction
  - sparseness of interaction
- TD-POMDP well-suited for **weakly-coupled problems** with sparse interactions



TD-POMDP Benefits

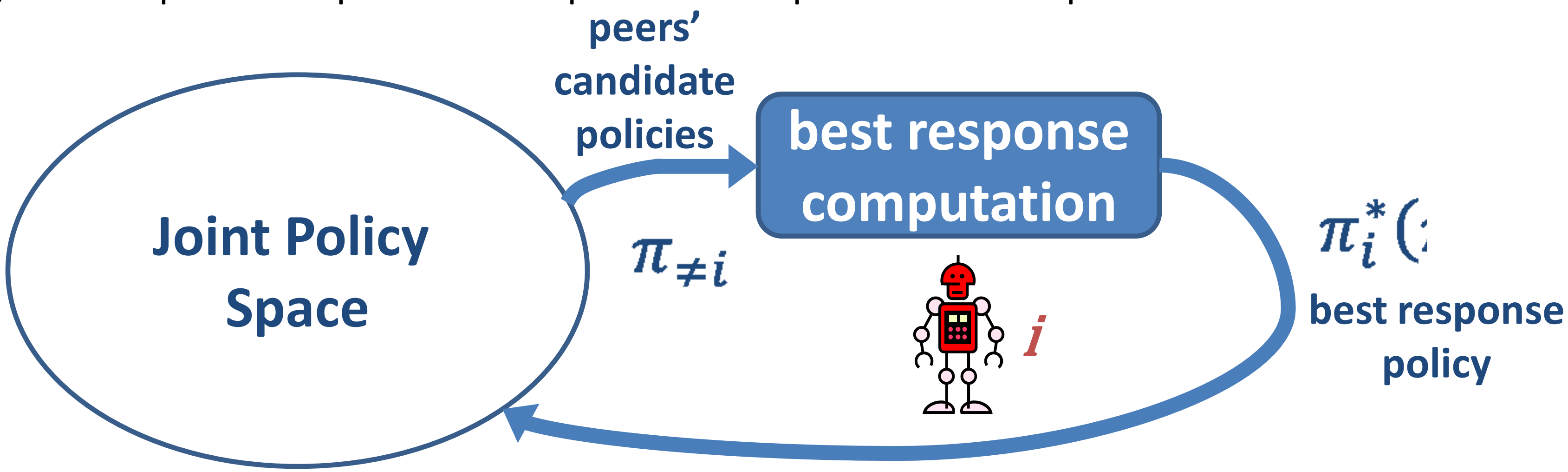
- Explicit representation of transition-dependent interaction features
- Naturally conveys
  - locality of interaction
  - sparseness of interaction
- TD-POMDP well-suited for **weakly-coupled problems** with sparse interactions



(interaction digraph)

Decoupled Solution Methodology

- best-response search through the joint policy space
  (e.g., JESP [Nair *et al*.], GOA [Nair *et al*.], CSA [Becker *et al*.], …)
- Agents compute local policies in response to the policies of their peers



→ Successful for scaling (transition & observation-independent) ND-POMDPs

→ Less so for transition-dependent Dec-POMDPs

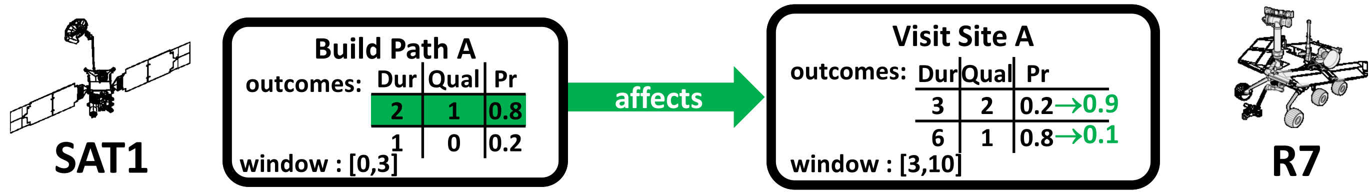
- Best-response model unwieldy
  - requires reasoning about other agents' possible observation histories
- Joint policy space very large

Best Response

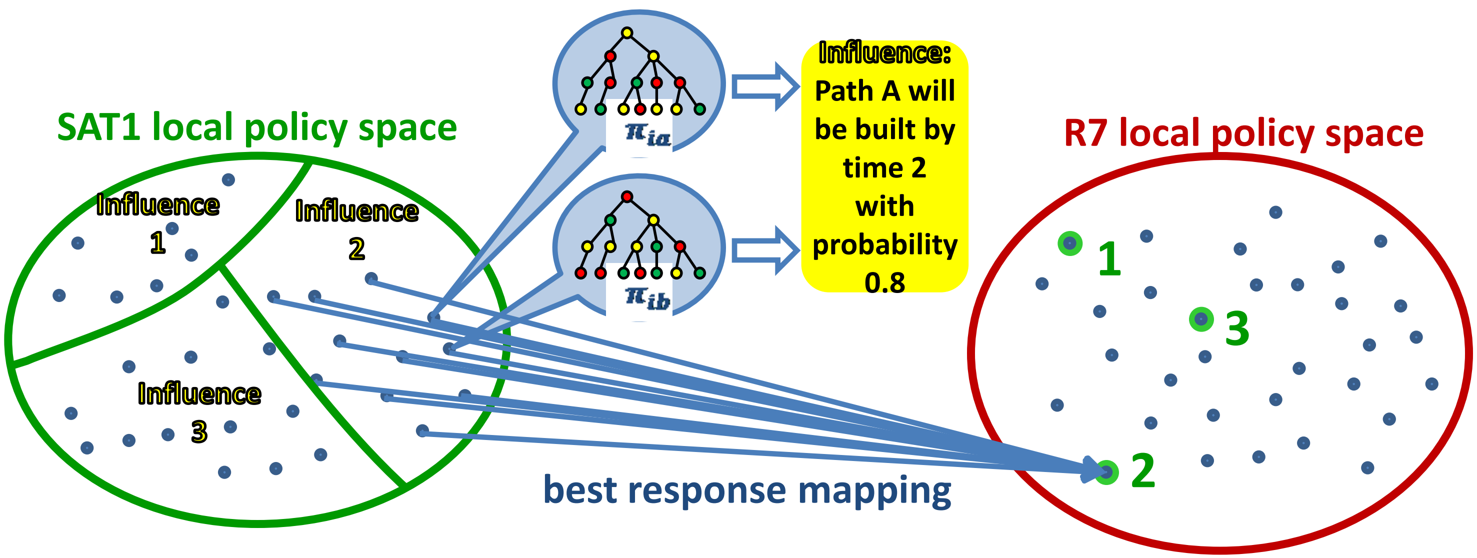
For a potential peer policy…

- Account for influence of peer's planned decisions on own decision-making problem
- Plan own decisions accordingly

Influence



- R7's behavior is only influenced by the likelihood of path A being built by time 3
- SAT1's decisions after time 3 have no impact on R7

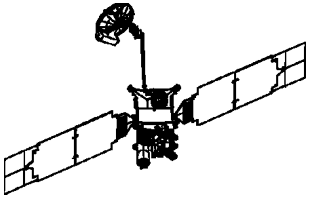


- For weakly-coupled problems…
  - Many peer policies map to the same influence
  - For all such policies, the best response will be the same!

Influence-based policy abstraction

…of *i*'s policy on j's decision-making problem

policy
π_i

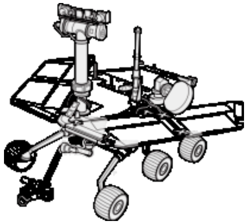


influence abstraction

influence
$\Gamma(\pi$

best response computation

$\pi_j^*($



TD-POMDP Influence Mechanics

- For TD-POMDP, the *influence* relates to the expected changes of nonlocal feature n_j

nonlocal features value

- Influence $\Gamma(\pi_i) = \{Pr\,(n_j \mid \cdots)\}$

values on which nonlocal feature value depends

Example: $Pr(path\text{-}A\text{-}built^{t+1} = T \mid path\text{-}A\text{-}built^{t} = F, t = 2) = \mathbf{0.8}$

TD-POMDP Influence Mechanics

policy π_i → **influence abstraction** → influence $\varGamma(\pi$

- For TD-POMDP, the *influence* relates to the expected changes of nonlocal feature n_j
- Influence $\varGamma(\pi_i) = \{Pr\,(n_j|\cdots)\}$

nonlocal features value

values on which nonlocal feature value depends

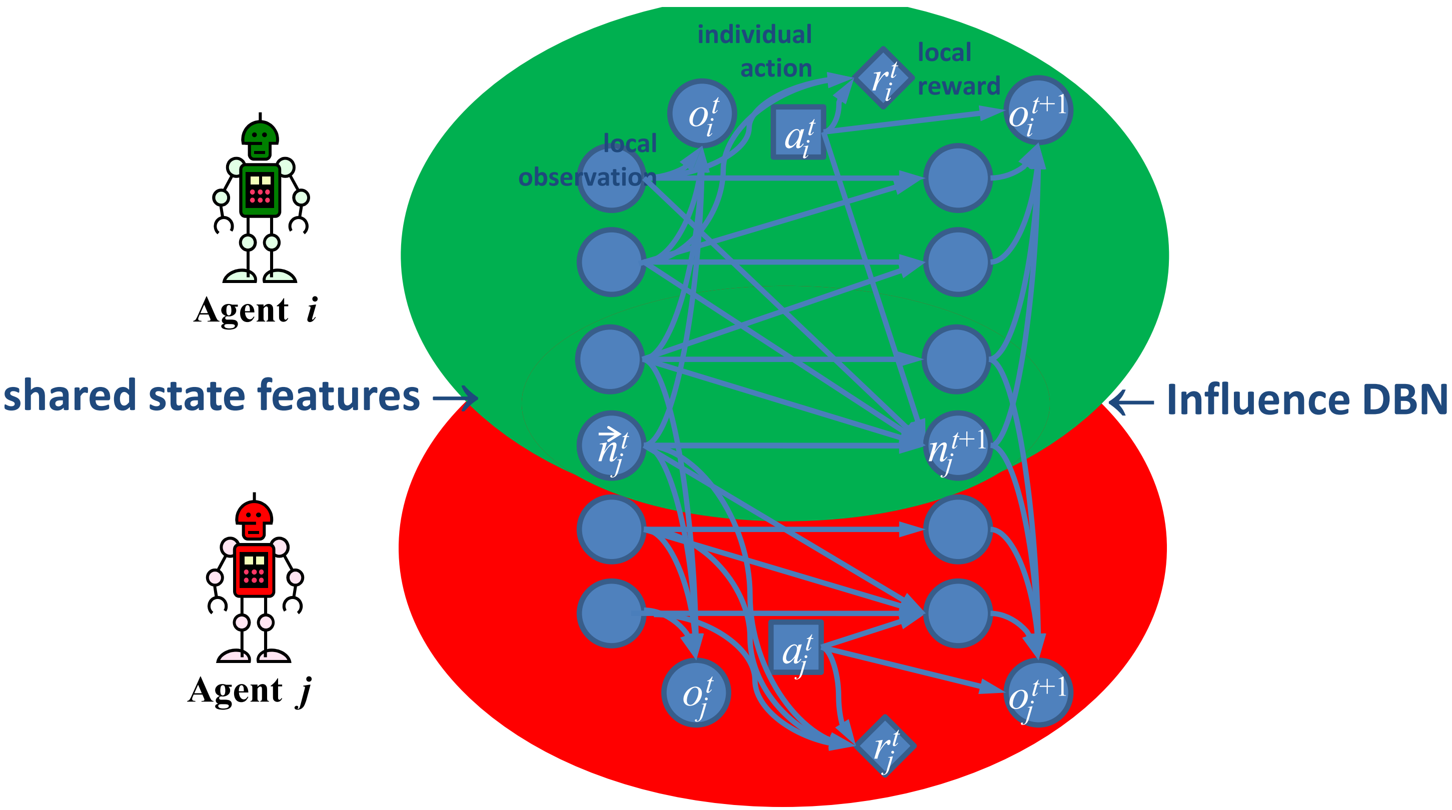
Example: $Pr(path\text{-}A\text{-}built^{t+1} = T | path\text{-}A\text{-}built^{t} = F, t = 2) = 0.8$

influence $\varGamma_i$ → **best response computation** → $\pi_j^*($ best-response policy

1) Create POMDP using TD-POMDP *local state* space, *local state* transitions, local observations, and local rewards
2) Augment state with variables on which influences depend
3) Set transitions of nonlocal features according to influence information

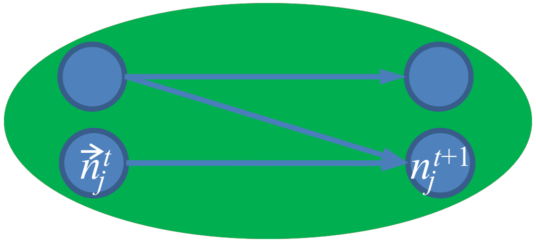
Sufficiency of Influence

- [Proposition 1] To compute consistent best responses, the influence distributions $Pr\left(n_j | \cdots\right)$ need only be conditioned on past and present values of shared state features



Sufficiency of Influence

- [Proposition 1] To compute consistent best responses, the influence distributions $Pr(n_j | \cdots)$ need only be conditioned on past and present values of shared state features



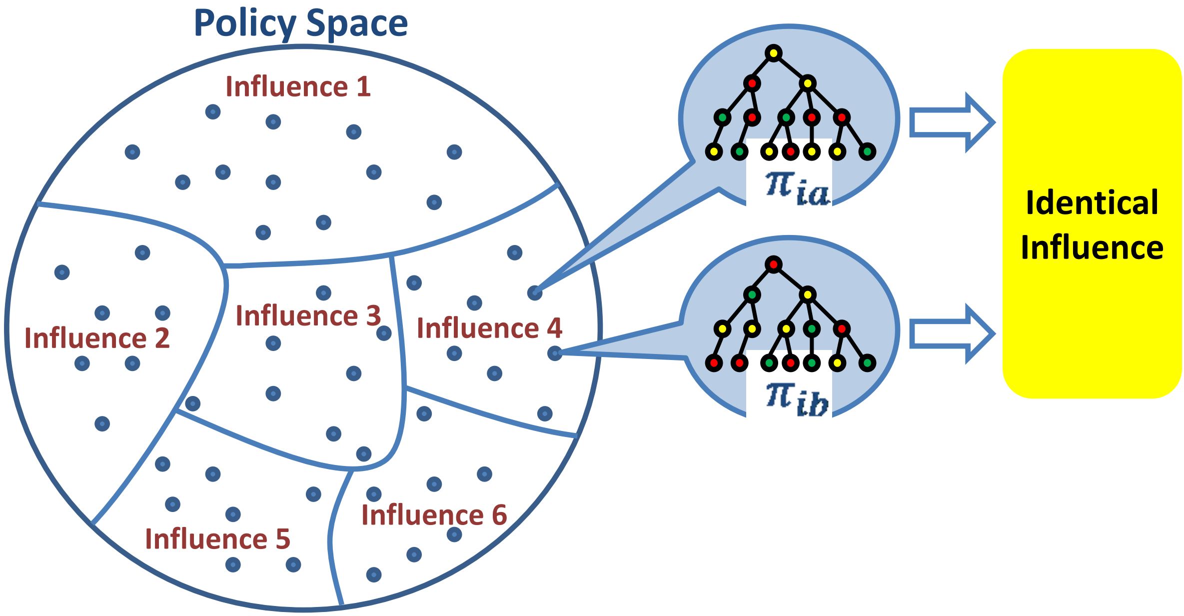
← Influence DBN

- For weakly-coupled TD-POMDP problems…

→ local *best-response* model compact

→ the number of parameters needed to represent influences remains small

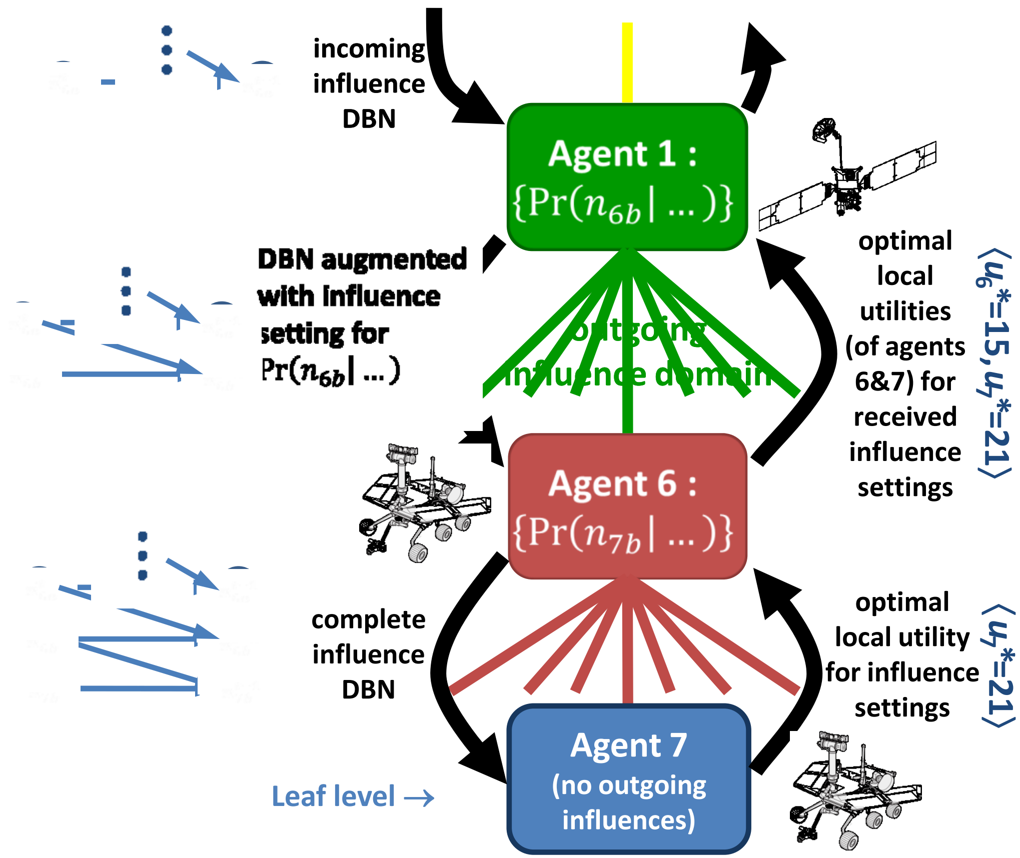
Influence Space



- Potentially significantly smaller than the policy space
- Optimal Influence $\rightarrow$ Optimal joint policy

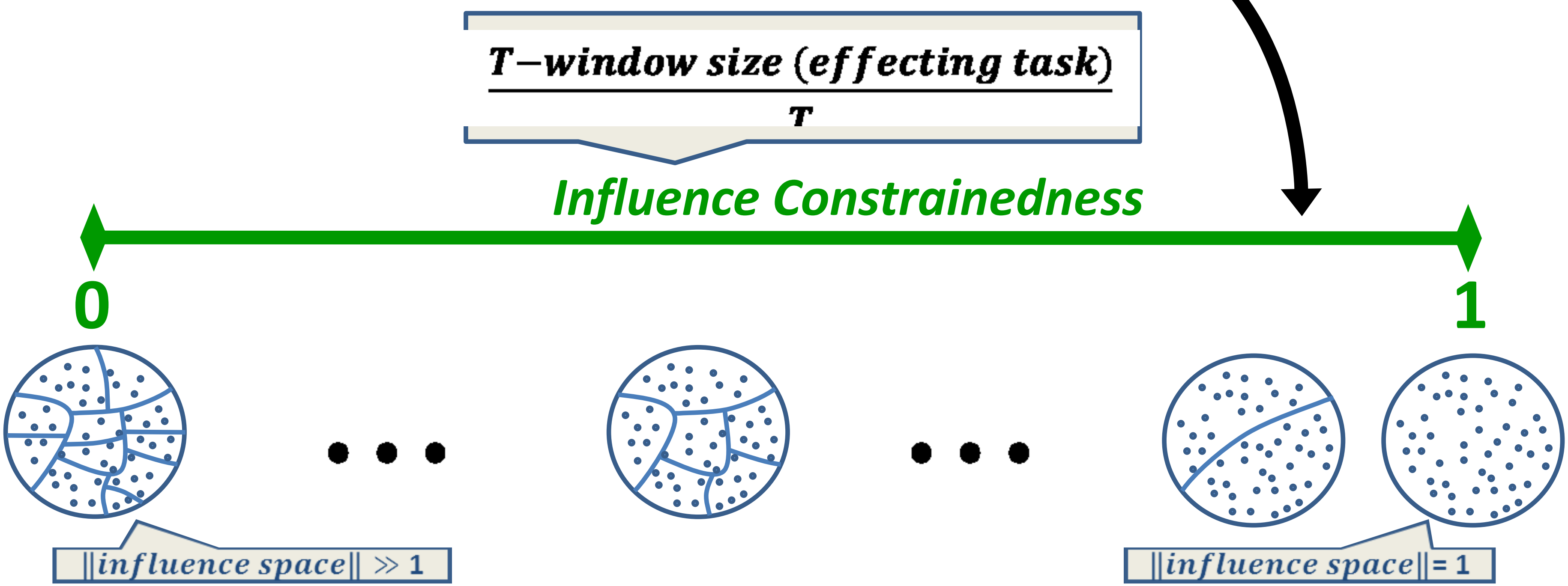
Optimal Influence-space Search **(OIS)**

- Depth-first search of agents' influence settings
  - Agents generate feasible influence settings and corresponding optimal local utilities (using Linear Programming)
  - Pass settings down
  - Pass local values back up



Hypothesis

- OIS has greatest advantage (over conventional policy-space coordination) on problems with…
  - Few interactions
  - Interactions which are highly constrained

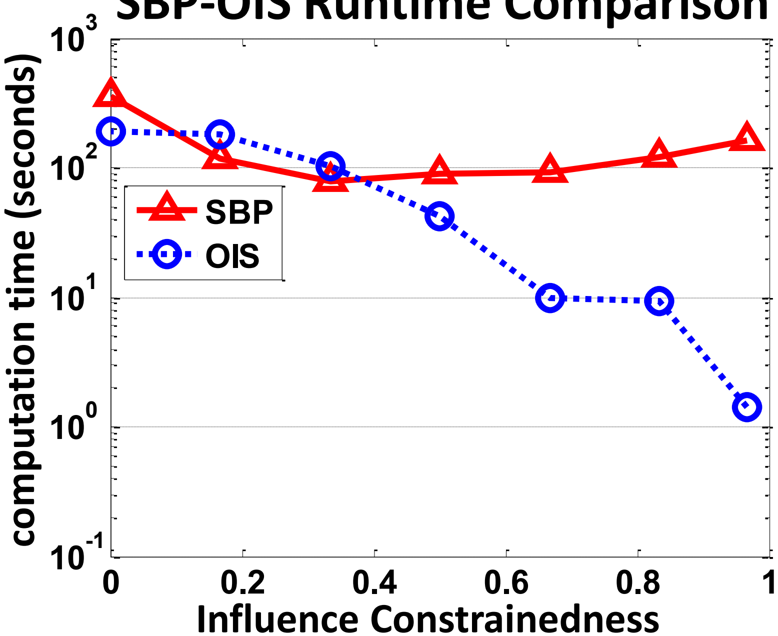


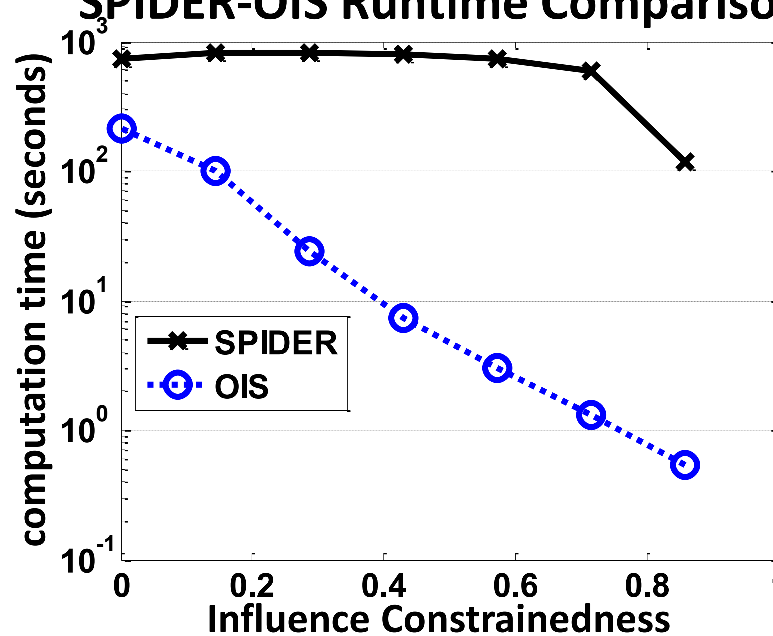
Empirical Comparison with Policy Search Methods

Separable Bilinear Programming [Mostafa *et al.*] : for EDI-CR (DEC-MDPs)

SPIDER, implementation for specialized transition-dependent Dec-POMDPs [Marecki, Varakantham *et al.*]

- Single nonlocal feature, dependent on shared time fea…





- 25 problems
- 2 agents
- 4 tasks each
- 2 outcomes per task
- T = 30 time units
- No “wait” action

$$x\text{-}axis: \frac{T - \text{window size (effecting task)}}{T}$$

- 25 problems
- 2 agents
- 3 tasks each
- 3 outcomes per task
- T = 7 time units

Hypothesis 2

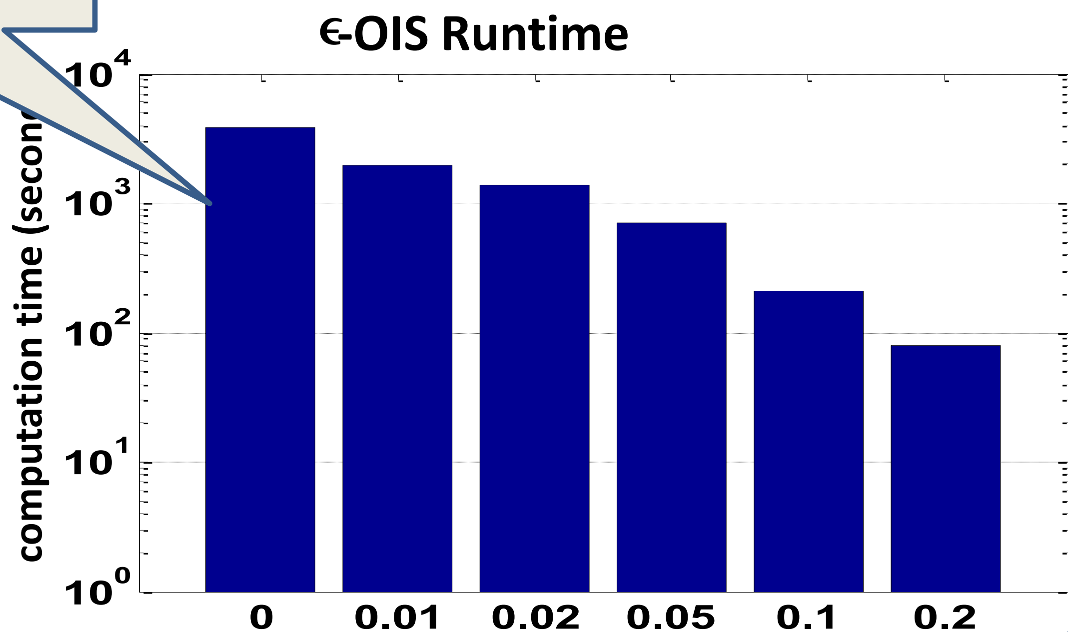
- Representation of influences using probability distributions enables flexible approximation

Strategy 1: only consider probability values that are $\geq \epsilon$ from those already found

Optimal solution for 4-agent transition-dependent Dec-POMDP problem!

ng and Approximation

- 10 problems
- 4 agents
- 3 tasks each
- 3 outcomes per task
- T = 6 time units
- Influence constrained-ness = 0.667



small loss of quality for exponential gain in efficiency.

ϵ	0	0.01	0.02	0.05	0.1	0.2
normalized joint utility value V	1.000	1.000	0.998	0.991	0.990	0.922
stddev(V)	0.000	0.000	0.005	0.020	0.020	0.117
improvement over uncoor-dinated local optimization	22.7 %	22.7 %	22.5 %	21.7 %	21.6 %	14.6 %
runtime	3832	1985	1369	702.2	208.9	79.80
stddev(runtime)	4440	2502	1995	1013	207.7	60.40

Flexible approximation!

Conclusions and Future Work

- Transition-Decoupled POMDP model
  - **General** planning model for weakly-coupled multi-agent system with **sparse transition-dependent interactions**
  - **Explicit representation** of interaction features
  - When peer policies are fixed, decouples into **compact optimal local (best-response) model**
- Influence-based Policy Abstraction
  - Influence space potentially significantly smaller than policy space (and no larger!)
  - No loss of solution quality (OIS guarantees optimal joint policy)
  - Agents need not exchange complete policies
  - Accommodates approximation flexibly
- Future Work
  - Empirical Evaluation on problems with varied agent coupling & interaction digraph structure
  - Empirical Comparison with approximate methods
  - Derivation of quality bounds for approximate versions of our algorithm

Thank You

- Questions?